scientific reports

OPEN

An innovative multi-head attention mechanism-driven recurrent neural network model with feature representation fusion for enhanced image captioning to assist individuals with visual impairments

Mashael M. Asiri[1]✉, Kholoud Alghamdi[2], Fahad Alzahrani[3] & Mahir Mohammed Sharif[4,5]

Developments in image captioning technologies played a crucial role in improving the quality of life for individuals with visual impairments, advancing better social inclusivity. Image captioning is the task of representing the visual content of the images in natural language, applying a language method and a visual understanding system able to generate significant and syntactically correct sentences. Image captioning is a field of research of vast significance, targeting the creation of natural language representations for visual content in static images. Automatically representing the image content is a significant challenge in artificial intelligence (AI). Therefore, the emergence of deep learning (DL) and the most recent vision-language pre-training methods have significantly advanced the domain, resulting in more advanced techniques and enhanced performance. DL-based methods can process the difficulties and nuances of image captioning. This paper proposes an Innovative Multi-Head Attention Mechanism-Driven Recurrent Neural Network with Feature Representation Fusion for Image Captioning Performance (MARNN-FRFICP) approach to assist individuals with visual impairments. The MARNN-FRFICP approach aims to enhance image captioning by employing an effective method focused on improving accessibility for individuals with visual impairments. Initially, the Gaussian filtering (GF) technique is utilized in the image pre-processing stage to enhance image quality by removing the noise. In addition, the fusion of advanced DL models, namely InceptionResNetV2, convolutional vision transformer (CvT), and DenseNetl69, is employed to enhance the effectiveness of the feature extraction process. Moreover, the hybrid of multi-head attention mechanism-based bi-directional long short-term memory and gated recurrent unit (MH-BLG) technique is used for classification. Finally, the Lyrebird optimization algorithm (LOA) technique is employed for tuning. The efficiency of the MARNN-FRFICP methodology is examined under the Flickr8k, Flickr30k, and MSCOCO datasets. The experimental analysis demonstrates that the MARNN-FRFICP methodology has improved scalability and performance compared to recent techniques in various measures.

Keywords Image captioning, Recurrent neural network, Attention mechanism, Feature representation fusion, Gaussian filtering, Visual disabilities

Vision and Language technology, specifically image captioning, can assist individuals with visual impairments in living more independent lives by describing the visual world around them in natural language. Image captioning has several applications, helping the visually impaired to understand images more effectively. Visually impaired

[1]Department of Computer Science, Applied College at Mahayil, King Khalid University, Abha, Saudi Arabia. [2]Department of Data Science, College of Computer Science and Engineering, Jeddah University, Jeddah, Saudi Arabia. [3]Department of Information and Computer Science, College of Computing and Mathematics, King Fahad University of Petroleum and Minerals, Dhahran, Saudi Arabia. [4]Department of Computer and Self Development, Preparatory Year Deanship, Prince Sattam bin Abdulaziz University, AlKharj, Saudi Arabia. [5]King Salman Centre for Disability Research, Riyadh 11614, Saudi Arabia. ✉email: abusharara@kku.edu.sa

individuals encounter difficulties when accessing visual information, which is essential for everyday life. Image captioning is a broad task in natural language processing (NLP) and computer vision (CV) that facilitates multimodal transformation from images to texts[1]. As a key provider of data, numerous images are stored and transmitted electronically over the internet. Simultaneously, social relationship mainly relies on NLP. It enables the processor to interpret the world of imagery. It has numerous applications, such as data recovery, support for visually impaired individuals, natural human-computer interaction, and education for children[2]. Depending on the input image, this method automatically generates a description of the text. As a demanding and significant AI domain, spontaneously created image descriptions have attracted substantial attention[3]. The aim is to create a verbal phrase that is linguistically precise in relation to the image's content. Therefore, visual understanding and language processing of image description are the dual main features of image captioning. The NLP and CV methodologies should be appropriately integrated to handle issues similar to those made by other models, ensuring the created sentence is both grammatically and semantically correct[4]. Automated image captioning is a significant research issue with various challenges, requiring a substantial workload with wide applications across multiple fields, including traffic data analysis, medical image captioning and prescription, human-computer interaction, quality industry control, and particularly assistive technology for visually impaired persons[5].

The area underwent innovative transformation through the growth and development of DL procedures, resulting in the introduction of progressive approaches and enhanced performance. Open-domain captioning is a highly challenging task, as it requires a detailed comprehension of both local and global elements in an image, along with their characteristics and relationships[6]. Image captioning is a well-explored area in AI, which involves understanding an image and generating descriptive text for it[7]. Image identification requests involve recognizing and detecting objects. Additionally, it aims to comprehend the location or scene type, its elements, and the relationships between them. Creating a well-formed sentence requires both semantic and syntactic understanding of language. Image understanding primarily relies on extracting features from the image. The methods employed for this motive are generally segmented into two types: (1) DL-based procedures and (2) Classical machine learning (ML)-based procedures. DL methodologies in sequence modelling have produced remarkable outcomes on the tasks, consistently leading the leaderboard[8]. Motivated by the newly presented decoder/encoder model for machine translation, which encodes the input image, the DL-based structures are trained end-to-end using back-propagation and achieve advanced results[9]. The use of spatial attention mechanisms for merging visual context—which indirectly indicates the generated text so far—was integrated into the generation procedure. It was shown that captioning methods employ attention mechanisms for optimal generality; thus, the DL comprise new text descriptions based on the detection of local and global objects[10].

This paper proposes an Innovative Multi-Head Attention Mechanism-Driven Recurrent Neural Network with Feature Representation Fusion for Image Captioning Performance (MARNN-FRFICP) approach to assist individuals with visual impairments. The MARNN-FRFICP approach aims to enhance image captioning by employing an effective method focused on improving accessibility for individuals with visual impairments. The efficiency of the MARNN-FRFICP methodology is examined under the Flickr8k, Flickr30k, and MSCOCO datasets. The key contribution of the MARNN-FRFICP methodology is listed below.

- The GF-based pre-processing is initially applied to remove noise and maintain spatial integrity in histopathological images, ensuring cleaner inputs for DL techniques, while improving feature clarity and consistency across the dataset. This step also enhances downstream processing and efficiently contributes to robust classification performance.
- The advanced fusion DL techniques, such as InceptionResNetV2, CvT, and DenseNet169, are employed for extracting rich, hierarchical, and multiscale features from histopathological images. This ability helps improve intrinsic patterns and subtle distinctions. This integrated feature representation significantly improves classification accuracy and robustness.
- The integrated LSTM and GRU model, namely hybrid MH-BLG, is implemented to improve sequential context learning and effectively capture temporal dependencies in feature sequences, thereby enabling more accurate and context-aware classification. This method enhances the interpretability and performance of the model across varied histopathological patterns.
- The LOA method is employed to fine-tune the model's parameters, thereby improving classification performance and convergence efficiency. It dynamically adjusts weights to prevent local optima and accelerates training, thereby assisting in robust decision-making for intrinsic histopathological image analysis.
- The novelty of the MARNN-FRFICP technique is in the integration of GF, InceptionResNetV2, CvT, and DenseNet169, and the MH-BLG classifier and LOA, within a single automated framework for histopathology. This incorporated model ensures robust noise handling, deep multiscale feature learning, and context-aware classification. The optimization also improves accuracy and convergence stability.

Related works on image captioning

Kalantari et al.[11] presented a method that integrates complicated semantic details with visual information to reconstruct. The suggested technique comprises dual components: semantic reconstruction and visual reconstruction. Visual details are encoded from brain data utilizing a decoder in the visual reconstruction model. The model uses a deep generator network (DGN) to produce images and employs VGG-19 models to extract visual characteristics from the generated images. Cao et al.[12] introduced the De-confounding Feature Fusion Transformer Network (DFFTNet) for image captioning, specifically intended to provide real-world support to visually impaired people. At the encoding stage, a distance-enhanced feature expansion method is used. This method efficiently develops the fine-grained information of image features by incorporating applicable positioning data within them. At the decoding stage, a causal adjustment model is projected to eliminate perplexing causes. Deepak et al.[13] developed a Residual Attention Generative Adversarial Network

(RAGAN) and utilized attention-based residual learning in GANs to enhance the fidelity and diversity of the generated image captions. The RAGAN leverages the word depending on the feature maps sooner to make good captions. The residual learning was implemented between the decoder and encoder networks. Al Badarneh et al.[14] examined transformer modules, highlighting the serious parts these attention mechanisms show. The projected module employs a transformer encoder–decoder design for generating text captions and utilizes a DL-CNN for image feature extraction. It presents a novel ensemble learning structure that enhances the quality of generated captions by utilizing multiple DNN frameworks that employ a voting mechanism. Padate and Kalla[15] introduced the Hybrid Chimp Wolf Pack Inception-V3 (HCWPI)-BiGRU model, which incorporates decoding and encoding components to generate precise captions for emotion-based input images. Firstly, the Chip Optimiser algorithm (COA) with the Wolf Pack Optimiser (WPO), such as HCWP, combined through InceptionV3 was used to generate proper reconstructions of the input image through fixed-length vectors, showing unique characteristics obtained during the encoding period. Lee et al.[16] presented an innovative method to generate scenes of AVs' safety. This method highlights the efficacy of the process in producing scenes by safeguarding representativeness and diversity. A multimodal image captioning module, denoted as Auto Scenario Generator (Auto-SG), is also employed, which spontaneously generates incidents using digital twin data. Arasi et al.[17] developed an Automatic Image Captioning employing the Sparrow Search Algorithm by the Improved DL (AIC-SSAIDL) method. The purpose of this AIC-SSAIDL procedure is to generate automatic text captions for input images. To achieve this, the AIC-SSAIDL procedure utilizes the MobileNet-V2 module to create input image feature descriptors, and its hyperparameter tuning practice is implemented using the SSA.

Deore et al.[18] introduced the Fully Convolutional Localisation Network (FCLN), a novel methodology that simultaneously addresses position and depiction challenges. It preserves spatial data and prevents information losses, restructuring training procedures by constant enhancement. The FCLN framework combines a recognition system, reminiscent of Faster R-CNN, with a caption method. This interaction allows us to produce image captions with semantic meaning. Safiya and Pandian[19] proposed a real-time image captioning system that utilizes visual geometry group 16 (VGG16) for feature extraction and long short-term memory (LSTM) for caption generation, deployed on a Raspberry Pi 4B. The NoIR camera is used for capturing images while also utilizing text-to-speech. Hossain, Anjom, and Chowdhury[20] introduced a lightweight U-Net-based model, named the Quantised Partial U-Net Lightweight Model (QPULM), and a Simple Obstacle Distance Detection (SODD) method. These are integrated into an Android app for real-time obstacle detection and footpath navigation using audio feedback. More et al.[21] developed an efficient object detection system using You Only Look Once version 8 (YOLOv8) converted to TensorFlow Lite for real-time assistance to visually impaired individuals. Anwar et al.[22] presented a computer-aided diagnosis system for glaucoma detection using an ensemble of ResNet-50, VGG-16, and Inception Version 3 (InceptionV3) models. Nguyen et al.[23] introduced MyUEVision, an Android-based assistive application. It also utilizes ExpansionNet V2 for online captioning and integrates VGG16 and LSTM for offline use. Qazi, Dewaji, and Khan[24] proposed a bilingual image captioning system by integrating a convolutional neural network (CNN) with a recurrent neural network (RNN), namely CNN-RNN, a vision transformer–generative pre-trained transformer 2 (ViT-GPT2), and generative adversarial networks (GANs), using translated Flickr30k captions. Muhammed Kunju et al.[25] presented a two-layer Transformer-based image captioning model by utilizing Inception V3 for feature extraction. This model is deployed on a Raspberry Pi 4B. Uikey et al.[26] presented a model by integrating Dense Convolutional Network 201 (DenseNet201) for feature extraction and LSTM for caption generation. These captions are converted into real-time audio using Google Text-to-Speech (gTTS), enhancing environmental awareness and independent mobility. Yousif and Al-Jammas[27] introduced a lightweight assistive system that utilizes YOLOv7 and a video Swin Transformer (ST) integrated with 2DCNN and Transformer networks for real-time object detection and video captioning on the Jetson Nano. Jenisha and Priyadharsini[28] developed a multi-layered image captioning system by utilizing CNN and LSTM networks to generate accurate captions for visually impaired users, which was enhanced with Google Text-to-Speech for audio output.

The limitations of existing studies include an insufficient exploration of vision foundation models, visual prompt tuning, few-shot adaptation, and neural architecture search, which are driving the current research frontiers. Additionally, various models still rely on conventional CNN-RNN techniques without integrating these advanced techniques, thereby restricting model adaptability and generalization. Furthermore, the robustness across diverse real-world scenarios remains a threat, and high computational costs also limit the deployment on edge devices. The research gap, when addressed, involves integrating these novel methods to enhance efficiency, flexibility, and performance in assistive image captioning systems for visually impaired users.

Methodological framework

In this manuscript, a new MARNN-FRFICP approach is proposed to assist individuals with visual impairments. The MARNN-FRFICP model aims to enhance image captioning through an effective approach focused on improving accessibility for individuals with visual impairments. It involves four processes: image pre-processing using the GF model, fusion of deep feature extraction models, and a hybrid MH-BLG model for automated image captioning. Figure 1 signifies the workflow of the MARNN-FRFICP model.

Pre-processing using GF

Initially, the GF technique is employed in the image pre-processing stage to enhance image quality by removing the noise[29]. This technique demonstrates excellence in mitigating noise while preserving crucial structural and spatial data in histopathological images. This model also maintains edge smoothness without introducing artefacts that could affect downstream analysis, unlike median or bilateral filters. The model is considered significant in medical imaging, where fine tissue patterns carry diagnostic relevance. This model also improves the clarity of the image, thus assisting DL techniques in extracting more meaningful features. The method

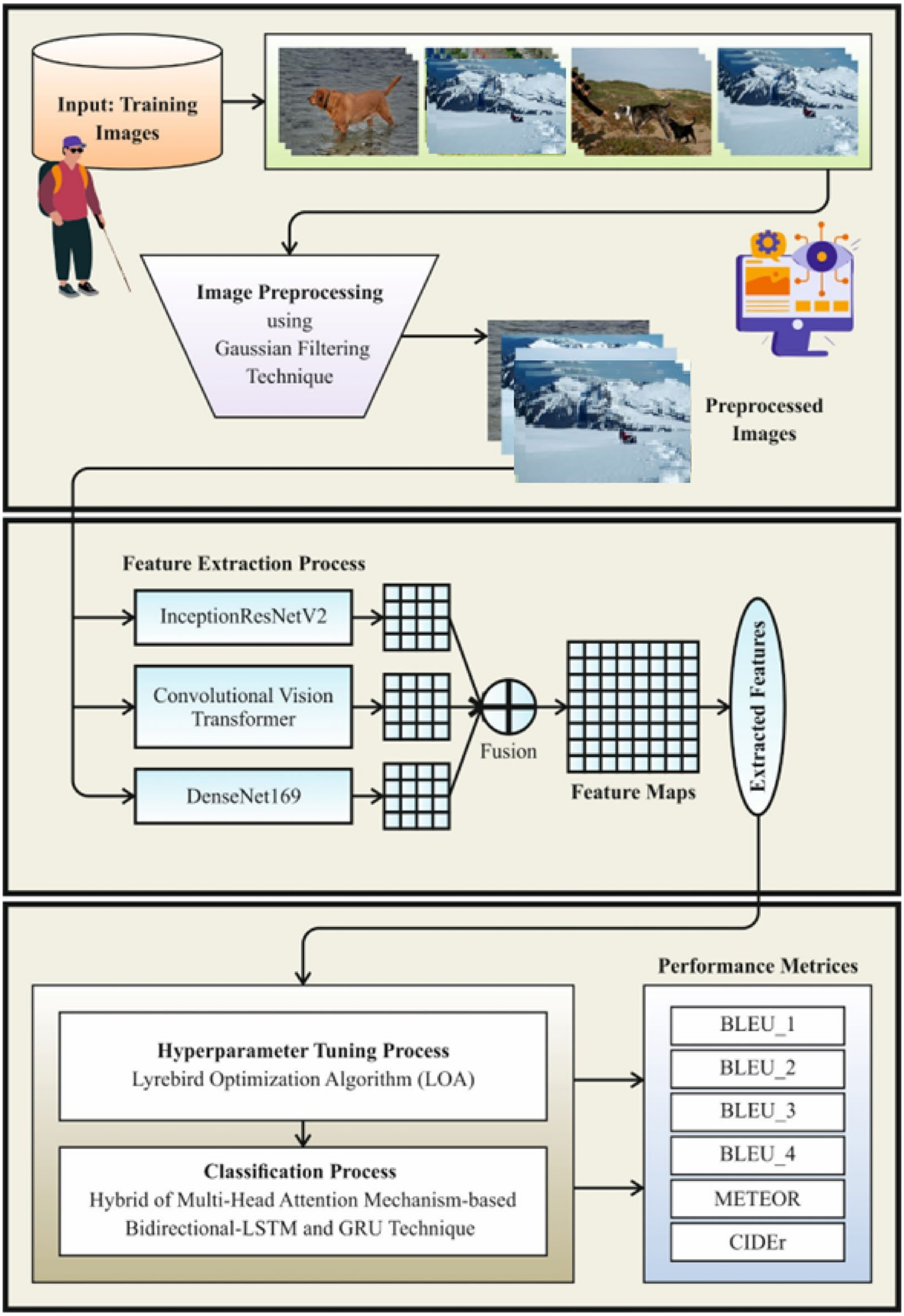
Input: Training Images
Image Preprocessing using Gaussian Filtering Technique
Preprocessed Images
Feature Extraction Process
InceptionResNetV2
Convolutional Vision Transformer
DenseNet169
Fusion
Feature Maps
Extracted Features
Performance Metrices
Hyperparameter Tuning Process
Lyrebird Optimization Algorithm (LOA)
Classification Process
Hybrid of Multi-Head Attention Mechanism-based Bidirectional-LSTM and GRU Technique
BLEU_1
BLEU_2
BLEU_3
BLEU_4
METEOR
CIDEr

Fig. 1. Workflow of the MARNN-FRFICP model.

effectively balances denoising performance and structural retention more effectively than various conventional filters, and its computational efficiency also makes it suitable for massive datasets.

GF is a widely employed model in image processing that reduces noise and smooths images while preserving the main structures. It utilizes a GF to weigh pixel values, thereby assigning high significance to adjacent pixels and reducing the impact of isolated ones. During image captioning, GF can improve the extraction of features by reducing unnecessary changes and enhancing object clarity. This enables DL techniques, such as transformers or CNNs, to focus on key image areas, resulting in more precise and contextually relevant captions. Moreover, its assistance in the pre-processing stages, by enhancing edge detection and reducing background noise, ultimately improves the model's performance in image captioning assignments.

Fusion of deep feature extraction models

In addition, the fusion of advanced three DL methods, namely InceptionResNetV2, CvT, and DenseNetl69 model, is employed to enhance the effectiveness of the feature extraction process. The fusion model is selected for its complementary strengths. The InceptionResNetV2 efficiently captures multiscale spatial features through inception modules integrated with residual connections, thus improving the depth without vanishing gradients. Furthermore, the CvT model incorporates local convolutional features and a global attention mechanism to facilitate better context understanding and fine-grained analysis. The DenseNet169 technique performs effective feature reuse through dense connectivity. This fusion model ensures robust, diverse, and hierarchical feature representation, while also mitigating redundancy and enhancing efficiency.

InceptionResNetV2

InceptionResNet-V2 is considered to be a multifaceted CNN structure mainly designed for operations in CV and image classification[30]. It continuously combines the succeeding dual important CNN models: Inception, known for its successful feature extraction, and ResNet, reliable for its capability to handle training difficulties in deeper systems. By combining the Inception module for ResNet's residual links and feature extraction, InceptionResNet-V2 exemplifies the power of both methods. Its structure consists of grid and stem elements. These units implement multiscale feature extraction, combining different convolutional methods while ensuring strong gradient flows over residual links.

The system begins with the input images, represented as a tensor, X_{input}, with dimensions $H\text{x}W\text{x}C$, where H and W represent the height and width of the images, and C refers to the channel count (for example, three for RGB images). These tensors are passed to the stem block, which serves as a pre-processing unit that removes lower-level features while minimizing spatial sizes.

$$X_{input} \in R^{H\times W\times C} \tag{1}$$

All convolution processes utilize kernels, $W_{i,c}$, to remove characteristics for all channels, c, of the input, followed by the inclusion of the biased term, b_j, and the application of the activation function, typically $ReLU$. In mathematics, the convolution process output for the ith feature mapping is stated as shown:

$$Z_i = ReLU\left(\sum_{c=1}^{c}(W_{i,c} * X_c) + b_i\right) \tag{2}$$

Whereas $*$ characterizes the convolution operator, Pooling tasks like average- or max-poolings are further related to decreasing the spatial sizes, as shown:

$$Z_{pool} = \max_{k\times k}(Z)\ or\ mean_{k\times k}(Z) \tag{3}$$

Here. $k\text{x}k$ refers to pooling kernel dimensions. The output of the stem block is the decreased spatial tensors, Z_{stem}.

It is a fundamental constituent part of the InceptionResNet-V2 structure, tailored to effectively remove composite features while preserving a smooth flow of gradients over residual links. This Inception module handles the input, X, over numerous equivalent divisions, all executing dissimilar processes. These branches contain the following:

1. 1 × 1 Convolutions to lower complexity and remove fine-grained attributes, as demonstrated:

$$Z_{1\times 1} = ReLU\left(W_{1\times 1}\times X + b_{1\times 1}\right) \tag{4}$$

2. 3 × 3 Convolutions using a reduction stage, whereas a 1 × 1 convolution decreases the channel counts previously used by a 3 × 3 convolution, as shown:

$$Z_{3\times 3} = ReLU\left(W_{3\times 3}\times\ \left(ReLU\left(W_r\times\ X + b_r\right)\right) + b_{3\times 3}\right) \tag{5}$$

3. 5 × 5 Convolutions separated into dual sequential 3 × 3 convolutions for computational cost, as represented:

$$Z_{5\times 5} = ReLU\left(W_{3\times 3b}\times\ \left(ReLU\left(W_{3\times 3b}\times\ X + b_{3\times 3b}\right)\right) + b_{3\times 3b}\right) \tag{6}$$

4. 1 × 1 convolutions accompany pooling, whereas pooling decreases the spatial sizes, and a 1 × 1 convolution is used for the compression of features, as indicated:

$$Z_{pool} = ReLU\left(W_{pool}\times\ \left(Pool\ \left(X\right)\right) + b_{pool}\right) \tag{7}$$

The corresponding branch outputs are connected along with the channel size, as illustrated:

$$Z_{Inception} = [Z_{1\times 1'}\ Z_{3\times 3'}\ Z_{5\times 5'}\ Z_{pool}] \tag{8}$$

The residual link expands the flow of gradient by including the input, X, to the Inception module outputs, as described:

$$Z_{\mathrm{Res}} = Z_{Incepiion} + W_r X \tag{9}$$

This ensures that the system can be trained to identify mappings that help alleviate the problem of gradient vanishing.

To mitigate memory usage and computing costs, reduction blocks are positioned among Inception Residual block groups. These blocks downsample the feature mapping over pooling operations and stride convolutions, as described:

I. Strided Convolutions:

$$Z_{reduce} = w_{reduce} \times\ X_{inpui} + b_{reduce} \tag{10}$$

II. Pooling additionally decreases spatial sizes, as specified:

$$Z_{reduce} = Pool\left(Z_{reduce}\right) \tag{11}$$

The block of reduction outputs a feature mapping, Z_{reduce}, with small spatial sizes and large channel counts. The last Inception Residual output blocks are restricted to a 1-D vector, as outlined:

$$z_{flatten} = Flatten\left(z_{final}\right) \tag{12}$$

Whereas $z_{final} \in\ R^{H_f \times\ W_f \times\ C_f}$ and the flattened vector $z_{flatten} \in\ R^{H_f \cdot W_f \cdot C_f}$.

This vector passes through more than one fully connected layer, as demonstrated:

$$Z_{dense} = W_{dense}.\ z_{final} + b_{dense} \tag{13}$$

Here, b_{dense} and W_{dense} represent the biases and weights of the dense layer. The last layers use a function of softmax activation for mapping the logit to likelihoods, as considered:

$$P\left(y_i|X\right) = \frac{\exp\left(Z_i\right)}{\Sigma_{j=1}^{N}\exp\left(Z_j\right)} \tag{14}$$

Now, Z_i denotes logit for class i, N means the sum of output class labels, and $P\left(y_i|\ X\right)$ stands for an anticipated possibility for class i. Figure 2 illustrates the architecture of InceptionResNetV2.

CvT model

CvT is a hybrid of a convolutional and Vision Transformer (ViT) framework[31]. The objective of this model is to leverage the benefits of CNNs in the fields of shared weights, local receptivity, and spatial subsampling, while incorporating the advantages of transformers in global context fusion, enhanced generalization, and dynamic attention.

$$x_{out} = \mathbb{T}\left(x_{in}{*}W_{in} + b_{in};\theta\ \right) \tag{15}$$

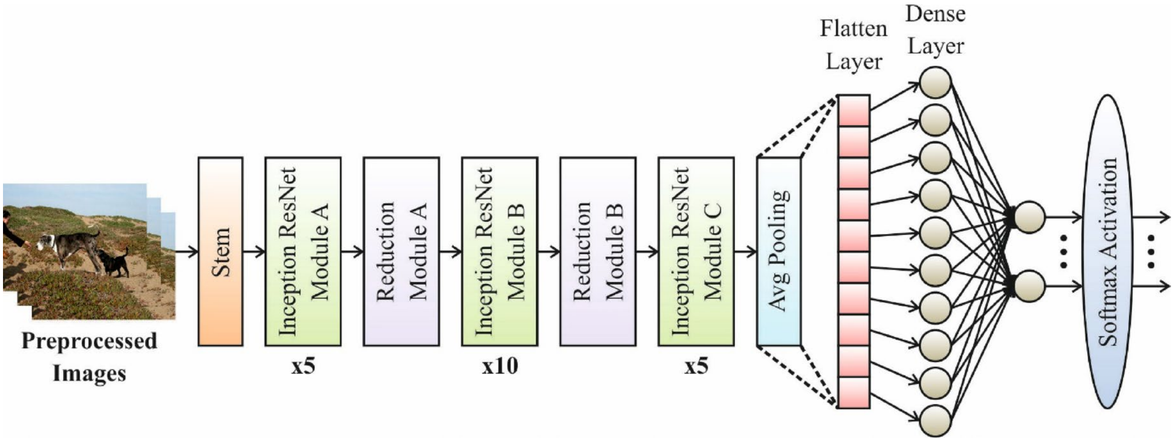

Fig. 2. Framework of InceptionResNetV2.

CvT framework might be expressed in Eq. (15), where x_{in} depicts features of the input, W_{in} and $b_{i}n$ refer to the weights and bias of the convolution embedding layer, and $\mathbb{T}\ (\bullet\,;\theta\ )$ denotes the transformer layer and parameters employed in that layer.

A transformer Block usually comprises dual elements. Primarily, the features are input and processed in the primary component like the normalization function and attention layers that are added to identity mapping of incoming features that is described in Eq. (16). Succeeding the 1 st portion, the features will be processed in the 2nd part that contains normalization function layer and Multi-layer Perceptron (MLP) layer, succeeded by identity mapping of features that enters 2nd part that stated in Eq. (17).

$$(x^{*} = x_{emb} + \mathbb{A}\,(\aleph\ (x_{emb})) \tag{16}$$

While the transformer processes the embedding outcomes, the feature passes through the layer of the attention mechanism, normalization, and creates a residual connection. In Eq. (16), $\aleph$ signifies a layer of normalization, x_{emb} indicates embedding outcome, x^{*} refers to normalization, attention, and residual connection results, and $\mathbb{A}$ represents the attention mechanism.

$$x_{out} = x^{*} + M\left(\aleph\ \left(x^{*}\right)\right) \tag{17}$$

In Eq. (16), the processing will be performed over the residual connection, normalization layer, and MLP. In Eq. (17), x^{*} depicts the calculation outcome of Eq. (16), $\aleph$ indicates the normalization layer, M denotes MLP calculated by Eq. (18), and x_{out} signifies the normalized MLP and residual connection outcomes.

$$x_{out} = \delta\ (x_{in}{*}W_1 + b_1)\,{*}W_2 + b_2 \tag{18}$$

Equation (18) is the equation of the MLP structure employed in Eq. (17). Within Eq. (18), x_{out} denotes the output feature, x_{in} indicates the input feature, W refers to the weight of the convolution layer, b depicts the bias of the convolution layer, and δ refers to the activation function.

DenseNetl69 method

The model's framework was advanced using DenseNet-169 for feature extraction[32]. By description, it is a mainstay, consisting of ConvNet layers two over 427, where all layers feed their feature mapping to the following layers, thereby encouraging feature reuse and the successful propagation of the gradient. This permits the dense connectivity form, which is highly effective in image processing for learning powerful features that are significant in differentiating subtle designs. As with the EfficientNet-B3 approach before, this mainstay is frozen to leverage the generalization control of its pre-trained weights on ImageNet. Features removed by the networks are accumulated over the layer of Global Average Pooling, resulting in a concise yet richer feature vector in terms of information. It further adds a fully connected (FC) dense layer utilizing 1024 units and ReLU for activation to improve higher-order representations, accompanied by a dropout layer to enforce some category of regularisation and prevent overfitting. This method concludes with an output layer of Softmax that forecasts likelihood distributions for both four different types and six types in the respective dataset. It will present efficient calculations with higher predictive precision due to its stronger yet basic structure; therefore, it is likely to be the best selection for the shown classification task.

Hybrid MH-BLG model for automated image captioning

Finally, the hybrid of the MH-BLG technique for the classification model[33]. This technique effectively captures both short- and long-range dependencies in sequential data, which is considered significant for evaluating the spatial and temporal patterns. The MH attention component enables focusing on various relevant features concurrently, thereby enhancing performance and interpretability. The Bi-LSTM efficiently captures bidirectional context, improving the comprehension of intrinsic image features, while GRU ensures efficient learning with mitigated computational cost. Thus, the incorporation of these components provides a balanced trade-off between accuracy and efficiency, outperforming conventional RNNs or single-layer LSTMs in handling high-dimensional medical image data.

The LSTM is an enhanced variant of the RNN, specifically designed to handle longer-term dependencies and time-series data. By presenting gating systems, LSTM dynamically regulates the retention and forgetting of data, efficiently resolving the problems of gradient explosion or vanishing that arise in conventional RNNs. This creates an LSTM well-matched to acquire data from longer-term dependencies. The framework of the LSTM network comprises three gates:

1. Forget gate (f_t): Establishes the data to discard from the state of the cell. The input consists of the existing input (X_t) and the preceding hidden state (HL) (h_{t-1}). Afterwards, passing over the sigmoid function, the value of the output is between 0 and 1. Here, 0 is a wide-ranging discard, and 1 refers to comprehensive retention.
2. Input gate (i_t): Regulates the upgrade of novel data. The sigmoid function's choice of values must be upgraded, as the $\tanh$ function creates a state of candidate $\left(\tilde{C}_t\right)$. The sigmoid outcome enhances the state of the candidate, and the outcome is added to the past state of the cell (C_{t-1}) to update the present state of the cell (C_t).
3. Output gate (o_t): Establish the HL of output. The existing cell state (C_t) is processed over the function of $\tanh$, and the outcomes are increased by the sigmoid function output by creating the present HL (h_t), thus

deliberating either the present input or longer-term memory. The mathematical model for the network of LSTM is given:

$$f_t = \sigma\ (W_f h_{t-1} + U_f X_t + b_f) \tag{19}$$

$$i_t = \sigma\ (W_i h_{t-1} + U_i X_t + b_i) \tag{20}$$

$$o_t = \sigma\ (W_0 h_{t-1} + U_o X_t + b_o) \tag{21}$$

$$\widetilde{C}_t = tanh\,(W_c h_{t-1} + U_c X_t + b_c) \tag{22}$$

$$C_t = f_t \times\ c_{t-1} + i_t \times\ \widetilde{C}_t \tag{22}$$

$$h_t = o_t \times\ tanh\,(C) \tag{24}$$

Here, $W_f,\ W_i,\ W_0,\ and\,W_c$ represent the weighted matrices for the state and gate units, $respectively.\ \sigma$ specifies the sigmoid activation function. These matrices are utilized to change the preceding HL $h_{t-1}\ linearly.$ $U_f,\ U_i,\ U_o,\ and\,U_c$ denote the weighted matrices for state and gate components. It is employed to change the existing input $X_t\ linearly.\ b_f,\ b_i,\ b_o,\ and\,b_c$ are the biased terms for the state and gate components.

The architecture of Bi-LSTM comprises both backwards and forward LSTM. It can process the input data in typical sequence order. For instance, the sequential data $X = (X_0,\ X_1,\ X_2,\ \ldots\ ,\ X_T)$ is processed by the LSTM in a forward direction, beginning from X_0 and determining the forward hidden state $\overrightarrow{h_t}$ at time step t. This process is then repeated for $X_1,\ X_2, \ldots\ ,\ X_T$ in sequence. Conversely, the reverse LSTM manages the input data in reverse, starting from X_T and analyzing the reverse HL $\overleftarrow{h_t}$ at time step t, then proceeding to $X_{T-1},\ X_{T-2}, \ldots\ ,\ X_0$. The Bi-LSTM output is used to concatenate the backwards and forward hidden states. Specifically, for every time-step t, the output is $Y_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$. The Bi-LSTM technique comprehensively deliberates either the preceding tendency or the upcoming tendency of sequential data, presenting more inclusive feature data for condition prediction and classification.

GRU has a version of the LSTM framework with a simple model. It integrates the forgetting and input gate of LSTM to a particular gate of upgrade and associates the cell states in the HL. Under similar concealed unit counts, the GRU method has a lower computation cost and a faster training speed compared to LSTM, particularly in processing large amounts of data. The architecture of the GRU model comprises dual gates:

Update gate (z_t): This gate regulates the number of the preceding memories that must be retained and the number of novel data that must be increased at the existing time step. The input in preceding HL (h_{t-1}) and the present input (X_t), after passing through the sigmoid function, creates an output value (z_t) between zero and one. While z_t is adjacent to one, it indicates that more preceding state data is retained; once it is near zero, it suggests that more reliable data is used on the current input to update the state.

Reset gate (r_t): This gate manages several preceding memories that must be retained. The input consists of the preceding HL (h_{t-1}) and the existing input (X_t). The output (r) has a value between zero and one; it is employed to manage the effect of the past hidden state (h_{t-1}) on the existing calculation. While the output of the gate of reset is zero, the past HL is disregarded; whereas it is one, the past hidden state is entirely deliberated.

The hidden state of candidate $\overrightarrow{h_t}$ is attained to concatenate and linearly change the scaled preceding HL h_{t-1} and existing input X_t. The existing HL h_t is measured by utilizing the output of the upgrade gate to manage the related contributions of the past hidden state h_{t-1} and the HL of the candidate $\widetilde{h}_t$. The gate of upgrade dynamically balances historical data and novel data, enabling the HL to acquire patterns and modifications in sequence data effectively. This method helps mitigate the problem of exploding and vanishing gradients when processing longer sequences, thereby enabling enhanced learning of longer-term dependencies. The mathematical model of GRU is given:

$$z_t = \sigma\ (W_z X_t + U_z h_{t-1} + b_z) \tag{25}$$

$$r_t = \sigma\ (W_r X_t + U_r h_{t-1} + b_r) \tag{26}$$

$$\widetilde{h}_t = \tanh\,(W_h X_t + U_h\,(r_t \times\ h_{t-1}) + b_h) \tag{27}$$

$$h_t = (1 - z_t) \times\ h_{t-1} + z_t \times\ \widetilde{h}_t \tag{28}$$

Now, $b_z,\ b_r,\ and\,b_h$ are the biased terms, $W_z,\ W_r,\ W_h,\ U_z,\ U_r,\ and\,U_h$ represent the weighted matrices, $and\,\sigma$ specifies the sigmoid activation function.

Attention mechanisms (AMs) are stimulated by human attention developments, which target the active concentration on different portions of input sequences during processing[34]. The MHAM extends the elementary AM by calculating numerous attention heads in parallel, permitting the method to concentrate on various attributes of the inputs. All attention heads learn diverse depictions that are further connected and linearly converted to give the last outputs:

$$Attention\ (Q,\ K,\ V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{29}$$

The outputs from numerous heads are concatenated:

Dataset	Images
Flickr8k	8000
Flickr30k	31,000
MSCOCO	164,062
Total	**203,062**

Table 1. Details of datasets.

Flickr8K Dataset						
Technique	**BLEU1**	**BLEU2**	**BLEU3**	**BLEU4**	**METEOR**	**CIDEr**
QPULM	60.04	44.93	34.20	20.05	16.32	32.96
YOLOv8	62.37	47.05	36.33	22.14	18.24	35.63
ResNet-50	64.47	48.41	38.09	24.87	19.98	38.33
Google NIC	60.00	44.86	34.12	19.99	16.23	32.89
Soft-Attention	62.29	46.97	36.29	22.07	18.19	35.58
m-RNN	64.38	48.33	38.05	24.81	19.91	38.26
SCA-CNN-VGG	66.98	51.81	41.01	26.28	23.23	40.10
GCN-LSTM	69.16	53.79	43.12	28.47	25.84	42.89
Injection-Tag	68.71	59.01	50.68	37.59	30.15	58.27
AIC-SSAIDL	74.19	58.19	47.69	33.48	31.53	47.95
MARNN-FRFICP	80.10	63.55	56.64	45.78	43.54	63.97

Table 2. Comparative analysis of MARNN-FRFICP model on Flickr8k dataset with recent models.

$$MultiHead\left(Q,K,V\right)=Concat\left(head_1,\ \dots\ ,head_h\right)W^O \tag{30}$$

Whereas Q, K, and V originate from the sequence of input features. Particularly, Q embodies the query gained by linearly converting the input characteristics; K signifies the key employed to calculate the similarities between Q and all input vectors; and V represents a matrix of values, which includes the real data of the input feature. W^O denotes a learnable output-weighted matrix that is constantly improved during training to integrate the outputs from dissimilar attention heads. These elements are automatically updated and calculated by the model during the input handling and training stages, eliminating the need for manual tasks. The layer of attention acts as a bridge between the output layer and the last hidden layers (HLs), particularly focusing on the more closely related features from the sequence of input. By allocating dissimilar weights to concealed states in HL, the AM enables the method to focus on significant spatial or temporal dependencies.

The combination of GRU, Bi-LSTM, and MHAM enhances the classification procedure in image captions by effectively capturing temporal and spatial dependencies. Bi-LSTM processes sequential image features bi-directionally, ensuring contextual understanding from both previous and upcoming states. GRU creates a recurrent network architecture that decreases computational complexity while retaining performance. The MHAM enhances feature representation by directing related portions of the image according to attention weights. This hybrid approach enhances caption generation precision, guaranteeing critical and contextually rich image representations.

Performance evaluation

The performance analysis of the MARNN-FRFICP model is examined under datasets such as Flickr8k[35], Flickr30k[36], and MSCOCO[37]. The complete details of these datasets are represented in Table 1.

Table 2; Fig. 3 present a comparative study of the MARNN-FRFICP technique on the Flickr8k dataset, alongside recent techniques, under several metrics, including BLEU1, BLEU2, BLEU3, BLEU4, METEOR, and CIDEr[17,20–22,38]. The table values specify that the MARNN-FRFICP technique has attained greater performance. Based on BLEU1, the MARNN-FRFICP technique has gained a maximum BLEU1 of 80.10%, while the existing methods, namely QPULM, YOLOv8, ResNet-50, Google NIC, Soft-Attention, m-RNN, SCA-CNN-VGG, GCN-LSTM, Injection-Tag, and AIC-SSAIDL, have reached a minimum BLEU1 of 60.04%, 62.37%, 64.47%, 60.00%, 62.29%, 64.38%, 66.98%, 69.16%, 68.71%, and 74.19%, correspondingly. Additionally, according to BLEU4, the MARNN-FRFICP technique achieved a superior BLEU4 score of 80.10%. In contrast, the existing methods, namely QPULM, YOLOv8, ResNet-50, Google NIC, Soft-Attention, m-RNN, SCA-CNN-VGG, GCN-LSTM, Injection-Tag, and AIC-SSAIDL, have achieved diminishing BLEU4 of 20.05%, 22.14%, 24.87%, 19.99%, 22.07%, 24.81%, 26.28%, 28.47%, 37.59%, and 33.48%, respectively. Additionally, according to METEOR, the MARNN-FRFICP technique has achieved a higher METEOR score of 43.54%. At the same time, the existing models, namely QPULM, YOLOv8, ResNet-50, Google NIC, Soft-Attention, m-RNN, SCA-CNN-VGG, GCN-LSTM, Injection-Tag, and AIC-SSAIDL, have achieved lower METEOR scores of 16.32%, 18.24%, 19.98%, 16.23%, 18.19%, 19.91%, 23.23%, 25.84%, 30.15%, and 31.53%, correspondingly. Furthermore, depending on CIDEr,

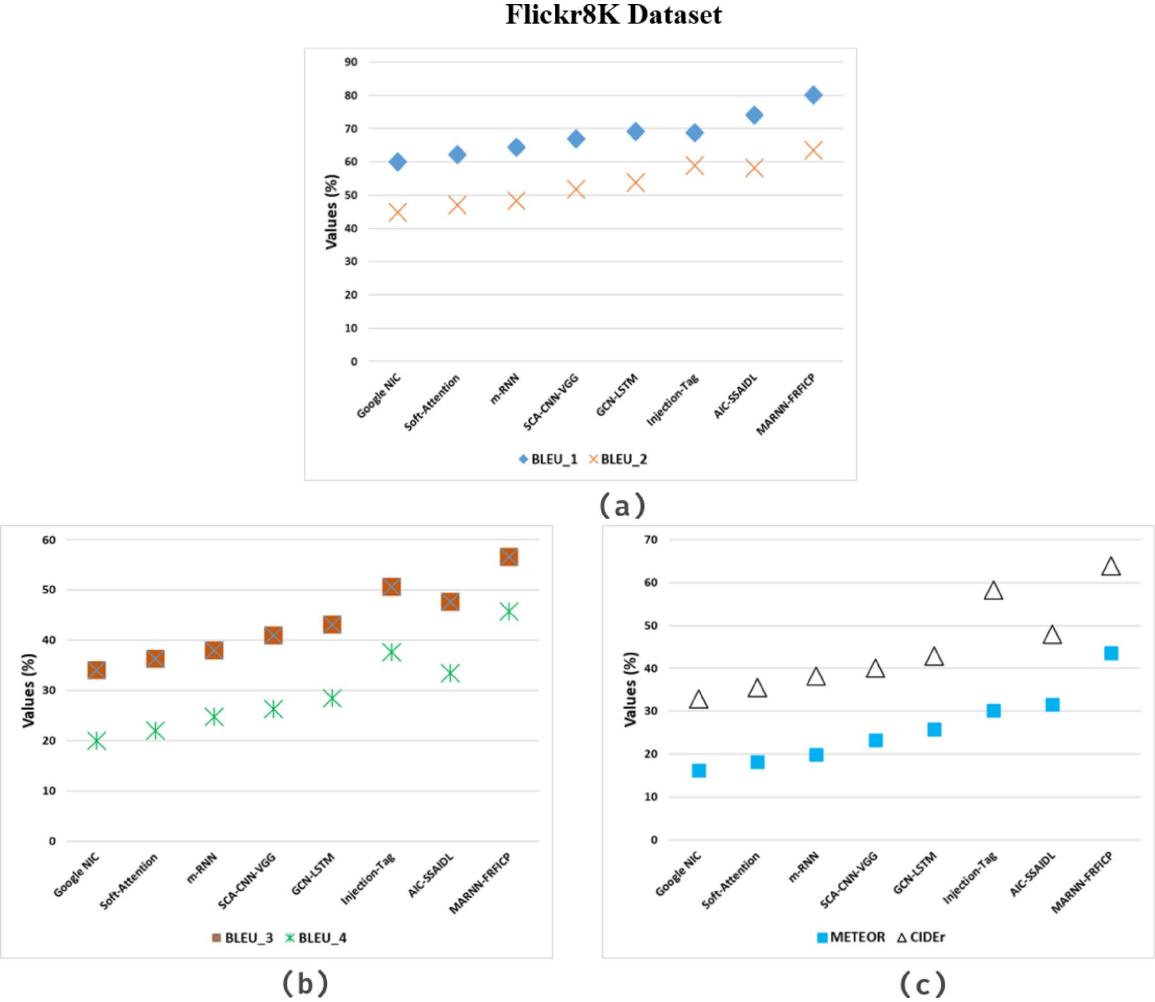

Fig. 3. Comparative analysis of the MARNN-FRFICP model on the Flickr8k dataset under various metrics.

the MARNN-FRFICP approach has achieved a higher CIDEr score of 43.54%. At the same time, the existing models, namely QPULM, YOLOv8, ResNet-50, Google NIC, Soft-Attention, m-RNN, SCA-CNN-VGG, GCN-LSTM, Injection-Tag, and AIC-SSAIDL, have achieved lower CIDEr scores of 32.96%, 35.63%, 38.33%, 32.89%, 35.58%, 38.26%, 40.10%, 42.89%, 58.27%, and 47.95%, respectively.

In Fig. 4, the training (TRA) $accu_y$ and validation (VAD) $accu_y$ performances of the MARNN-FRFICP method on the Flickr8k dataset are depicted. The figure underscored that both $accu_y$ values express a cumulative propensity, indicating the capability of the MARNN-FRFICP approach to achieve higher outcomes through numerous repetitions. Moreover, both $accu_y$ and results improve over time through the epochs, indicating diminished overfitting and presenting an increased outcome of the MARNN-FRFICP approach, which guarantees consistent prediction on unseen samples.

In Fig. 5, the TRA loss (TRALS) and VAD loss (VADLS) graph of the MARNN-FRFICP technique on the Flickr8k dataset is showcased. It is demonstrated that both values represent a declining propensity, indicating the proficiency of the MARNN-FRFICP approach in harmonizing a trade-off between generalization and data fitting. The constant reduction in loss values, as well as securities, provides an increased outcome from the MARNN-FRFICP approach, which ultimately tunes the calculation results over time.

The computational time analysis of the MARNN-FRFICP technique under the Flickr8k dataset is illustrated in Table 3; Fig. 6. The MARNN-FRFICP technique achieved the lowest CT of 3.77 s, significantly outperforming existing techniques such as GCN-LSTM with 21.61 s, Injection-Tag with 23.92 s, and Soft-Attention with 19.29 s. Compared to faster models like SCA-CNN-VGG with 6.67 s and m-RNN with 8.08 s, the MARNN-FRFICP model demonstrates superior processing speed, highlighting its computational efficiency and suitability for real-time or resource-constrained image captioning tasks.

Table 4; Fig. 7 depict the highest scores across all evaluation metrics, including BLEU-1 to BLEU-4, METEOR, and CIDEr of the MARNN-FRFICP model on the Flickr8K dataset. It attained a BLEU-4 score of 45.78, METEOR of 43.54, and CIDEr of 63.97, outperforming recent methods like MH-BLG and LOA. These results highlight the efficiency of the MARNN-FRFICP method in generating accurate and contextually rich image captions. The consistent performance gains also support its contribution in ablation studies, emphasizing the value of each integrated component.

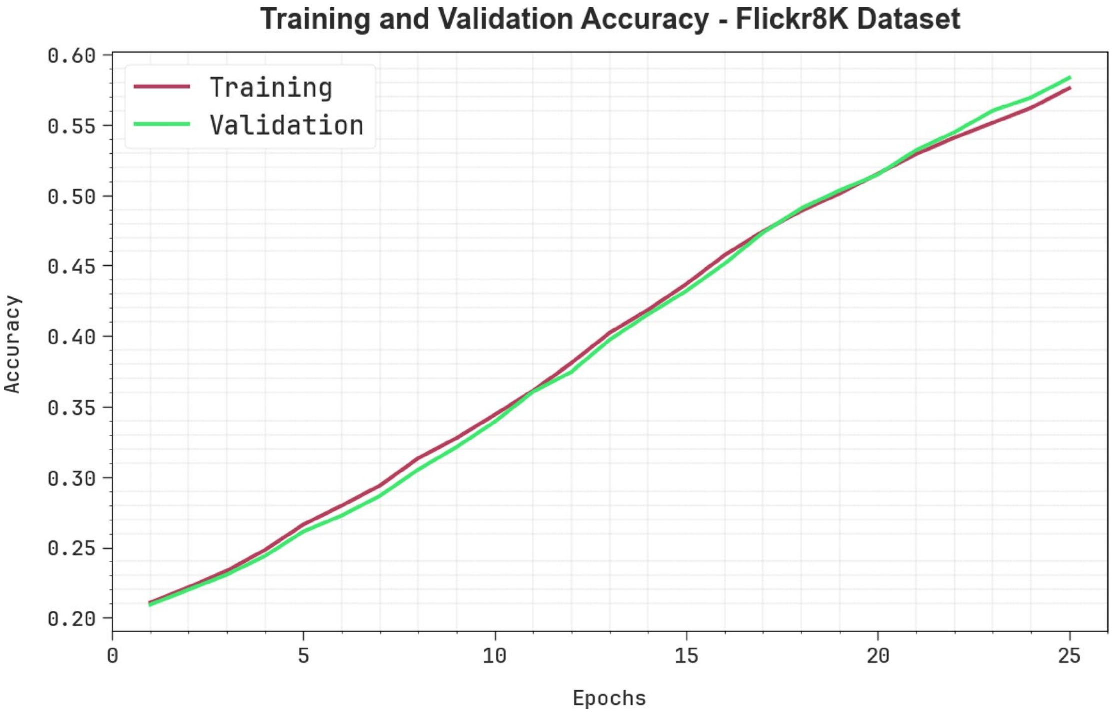

Fig. 4. $Accu_y$ curve of MARNN-FRFICP model on Flickr8k dataset.

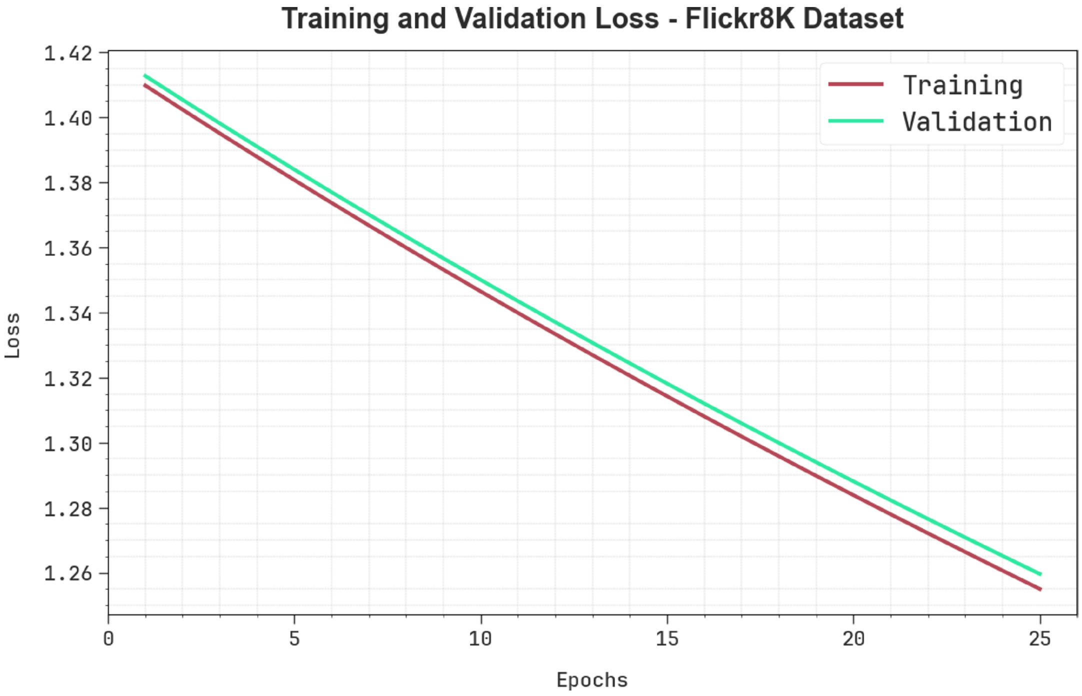

Fig. 5. Loss curve of the MARNN-FRFICP model on the Flickr8k dataset.

Table 5; Fig. 8 inspect the comparative study of the MARNN-FRFICP model with existing techniques under various metrics on the Flickr30K dataset. The outcome reports that the MARNN-FRFICP model has attained superior values in BLEU1 at 77.23%, BLEU2 at 70.11%, BLEU3 at 69.08%, and BLEU4 at 58.91%. In the meantime, the existing methods, such as QPULM, YOLOv8, ResNet-50, Google NIC, Soft-Attention, m-RNN, SCA-CNN-VGG, GCN-LSTM, Injection-Tag, and AIC-SSAIDL, have gained minimal values. Likewise, the AIC-SSAIDL method has provided More accurate solutions, with METEOR scores of 36.86% and CIDEr scores of 63.42%. In

Flickr8K Dataset	
Technique	**CT (sec)**
QPULM	15.99
YOLOv8	10.20
ResNet-50	11.98
Google NIC	13.08
Soft-Attention	19.29
m-RNN	8.08
SCA-CNN-VGG	6.67
GCN-LSTM	21.61
Injection-Tag	23.92
AIC-SSAIDL	12.65
MARNN-FRFICP	3.77

Table 3. CT evaluation of MARNN-FRFICP technique on Flickr8k dataset with recent models.

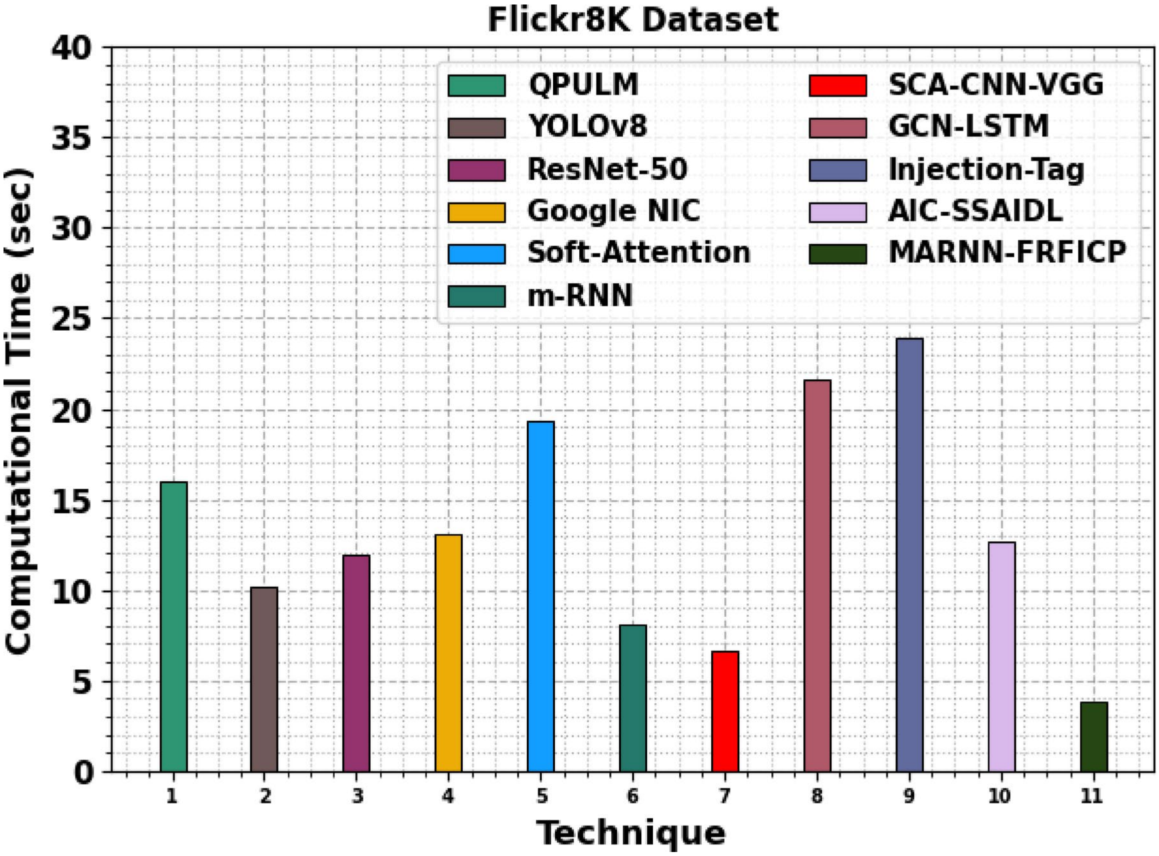

Fig. 6. CT evaluation of MARNN-FRFICP technique on Flickr8k dataset with recent models.

addition, the MARNN-FRFICP technique demonstrated maximum performance with an increased METEOR of 45.26% and a CIDEr score of 69.81%.

In Fig. 9, the TRA $accu_y$ and VAD $accu_y$ performances of the MARNN-FRFICP approach on the Flickr30K dataset are depicted. The figure emphasized that both $accu_y$ values present an increasing propensity, indicating the proficiency of the MARNN-FRFICP model with enhanced outcomes through several repetitions. In addition, both $accu_y$ remain closer through the epochs, indicating diminished overfitting, and exhibit the higher outcome of the MARNN-FRFICP model, securing consistent calculation on hidden samples.

In Fig. 10, the TRALS and VADLS of the MARNN-FRFICP approach on the Flickr30K dataset are exposed. It is indicated that both values exhibit a diminishing tendency, suggesting the competency of the MARNN-FRFICP approach in balancing the trade-off between generalization and data fitting. The constant reduction in loss values, along with assurances of improved outcomes for the MARNN-FRFICP technique, and gradual tuning of the prediction results, is a significant development.

Flickr8K Dataset						
Technique	**BLEU_1**	**BLEU_2**	**BLEU_3**	**BLEU_4**	**METEOR**	**CIDEr**
GF	76.33	59.43	52.31	41.38	39.13	59.82
InceptionResNetV2	76.84	60.20	53.16	42.11	40.00	60.37
CvT	77.45	60.90	53.72	42.78	40.61	61.10
DenseNet169	78.28	61.52	54.53	43.54	41.14	61.99
LOA	78.82	62.07	55.39	44.16	41.94	62.63
MH-BLG	79.47	62.73	56.13	45.00	42.74	63.25
MARNN-FRFICP	80.10	63.55	56.64	45.78	43.54	63.97

Table 4. Ablation study of the MARNN-FRFICP methodology on Flickr8k dataset.

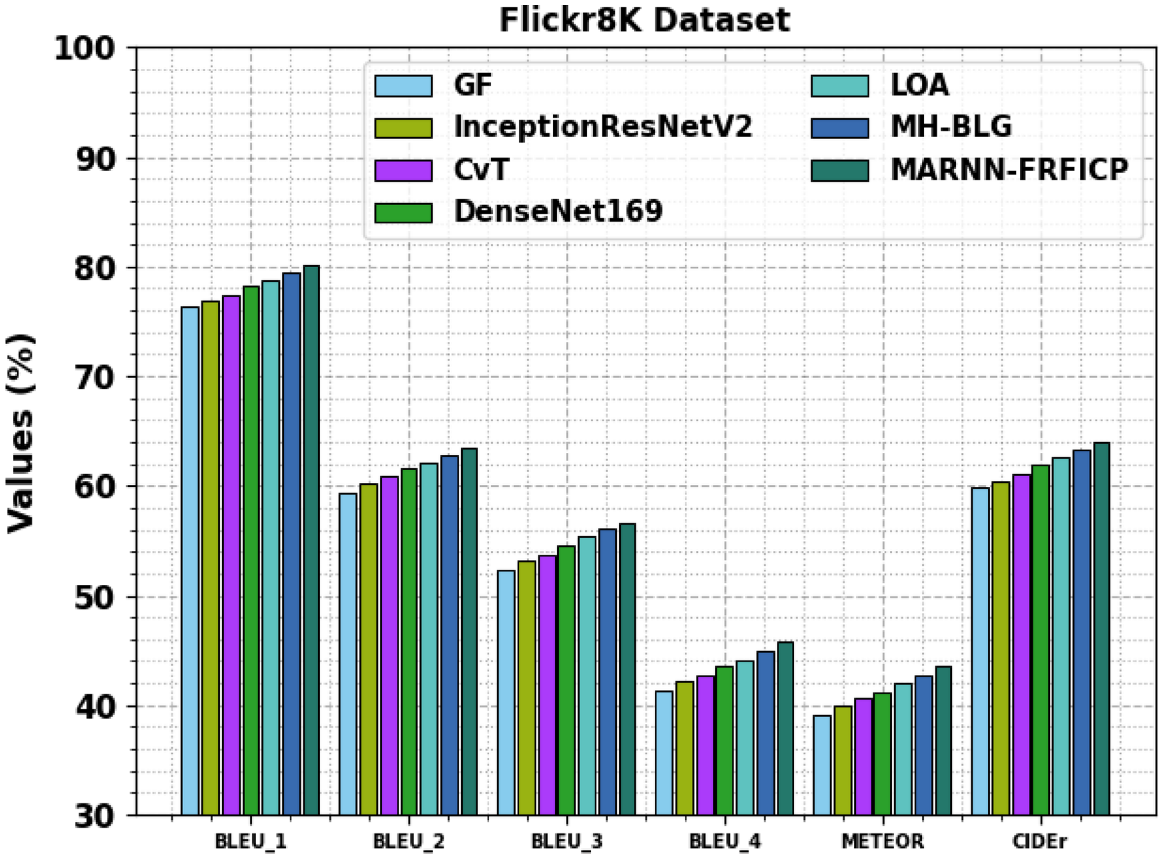

Fig. 7. Ablation study of the MARNN-FRFICP methodology on Flickr8k dataset.

Table 6; Fig. 11 specify the CT outputs of the MARNN-FRFICP technique under the Flickr30K dataset. While the MARNN-FRFICP model attained the lower value of 6.38 s, the existing models like Injection-Tag at 9.43 s, ResNet-50 at 12.19 s, and YOLOv8 at 23.70 s attained higher results. These values highlight the computational efficiency of the MARNN-FRFICP model, making it appropriate for real-time image captioning applications on larger and more diverse datasets like Flickr30K.

Table 7; Fig. 12 indicate the ablation study analysis of the MARNN-FRFICP technique on the Flickr30K dataset. The MARNN-FRFICP technique achieved the highest performance across all evaluation metrics, including BLEU-1 to BLEU-4, METEOR, and CIDEr. It recorded a BLEU-4 score of 58.91, METEOR of 45.26, and CIDEr of 69.81, surpassing robust baselines like MH-BLG and LOA. These improvements indicate that the MARNN-FRFICP model provides more fluent and semantically accurate captions. The consistent gains also reinforce the value of its individual components, as further validated through ablation studies.

The comparative outcome of the MARNN-FRFICP technique with existing methodologies under various metrics on the MSCOCO dataset is portrayed in Table 8; Fig. 13. Based on BLEU1, the MARNN-FRFICP technique has gained a maximum BLEU1 of 83.36%, while the existing models, namely QPULM, YOLOv8, ResNet-50, Google NIC, Soft-Attention, m-RNN, SCA-CNN-VGG, GCN-LSTM, Injection-Tag, and AIC-SSAIDL, have attained a minimum BLEU1 of 63.18%, 65.34%, 67.55%, 63.12%, 65.27%, 67.49%, 68.21%, 69.95%,

Flickr30K Dataset						
Technique	**BLEU1**	**BLEU2**	**BLEU3**	**BLEU4**	**METEOR**	**CIDEr**
QPULM	60.49	50.82	40.30	29.14	23.46	39.58
YOLOv8	62.91	53.21	43.16	30.77	25.67	41.68
ResNet-50	64.61	55.02	46.10	33.08	27.55	44.17
Google NIC	60.45	50.74	40.23	29.09	23.39	39.51
Soft-Attention	62.82	53.14	43.12	30.73	25.60	41.59
m-RNN	64.55	54.94	46.02	33.01	27.47	44.13
SCA-CNN-VGG	66.45	57.71	49.01	35.90	30.24	47.04
GCN-LSTM	69.22	59.28	50.74	37.74	30.74	58.49
Injection-Tag	71.00	61.97	52.56	39.69	34.47	61.17
AIC-SSAIDL	73.21	64.35	55.47	42.82	36.86	63.42
MARNN-FRFICP	77.23	70.11	69.08	58.91	45.26	69.81

Table 5. Comparative study of MARNN-FRFICP model with existing techniques on Flickr30K dataset.

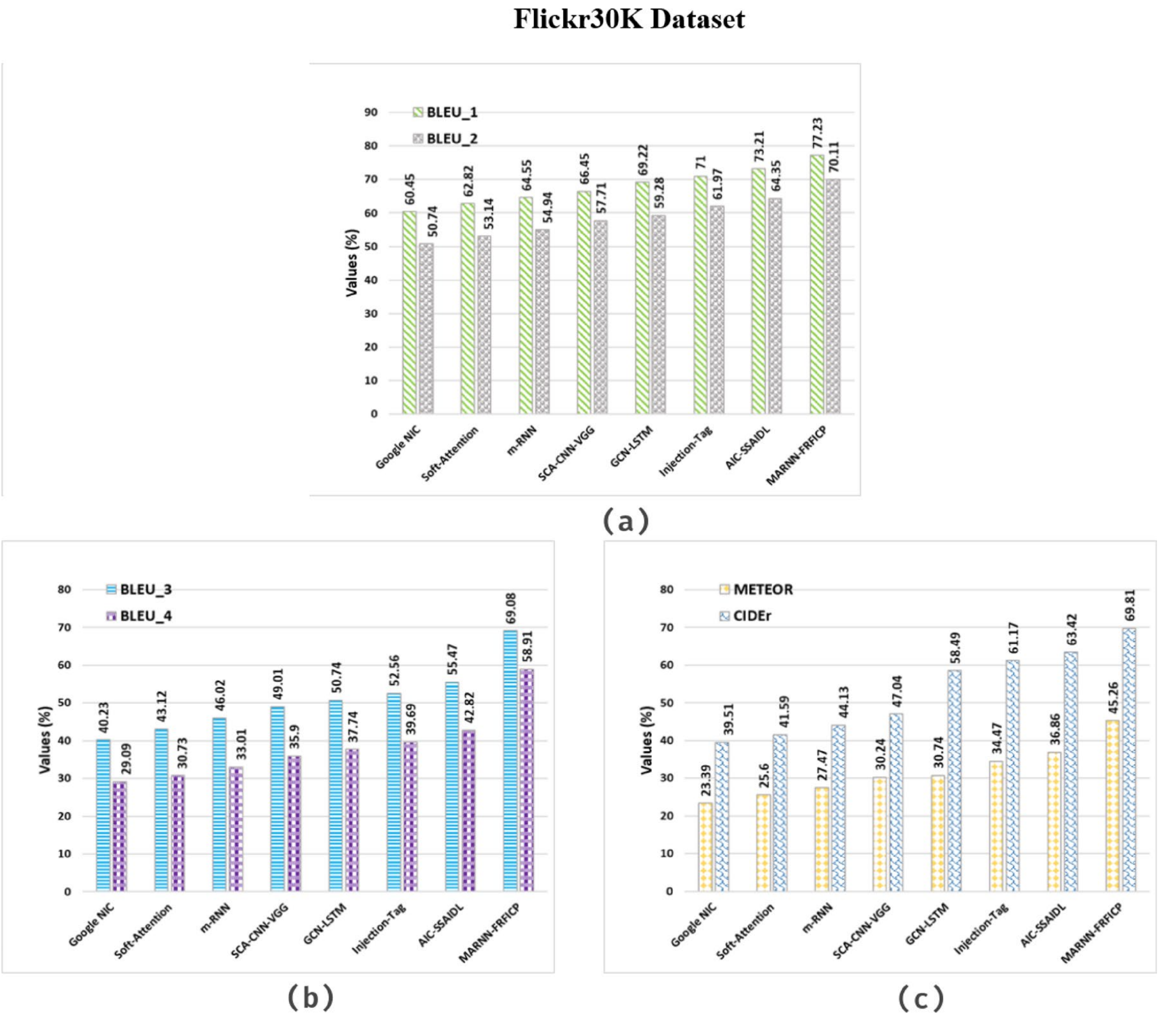

Fig. 8. Comparative analysis of the MARNN-FRFICP model on the Flickr30K dataset under various metrics.

76.73%, and 81.14%, respectively. Moreover, based on BLEU3, the MARNN-FRFICP technique has achieved a maximum BLEU3 of 83.36%. At the same time, the existing models, namely QPULM, YOLOv8, ResNet-50, Google NIC, Soft-Attention, m-RNN, SCA-CNN-VGG, GCN-LSTM, Injection-Tag, and AIC-SSAIDL, have achieved a minimum BLEU3 of 34.22%, 36.22%, 38.17%, 34.14%, 36.16%, 38.09%, 38.86%, 40.82%, 43.29%, and 48.59%, respectively. In addition, based on METEOR and CIDEr, the performance reports that the MARNN-FRFICP method has attained higher values in METEOR of 41.69%, and CIDEr of 150.62% whereas the recent

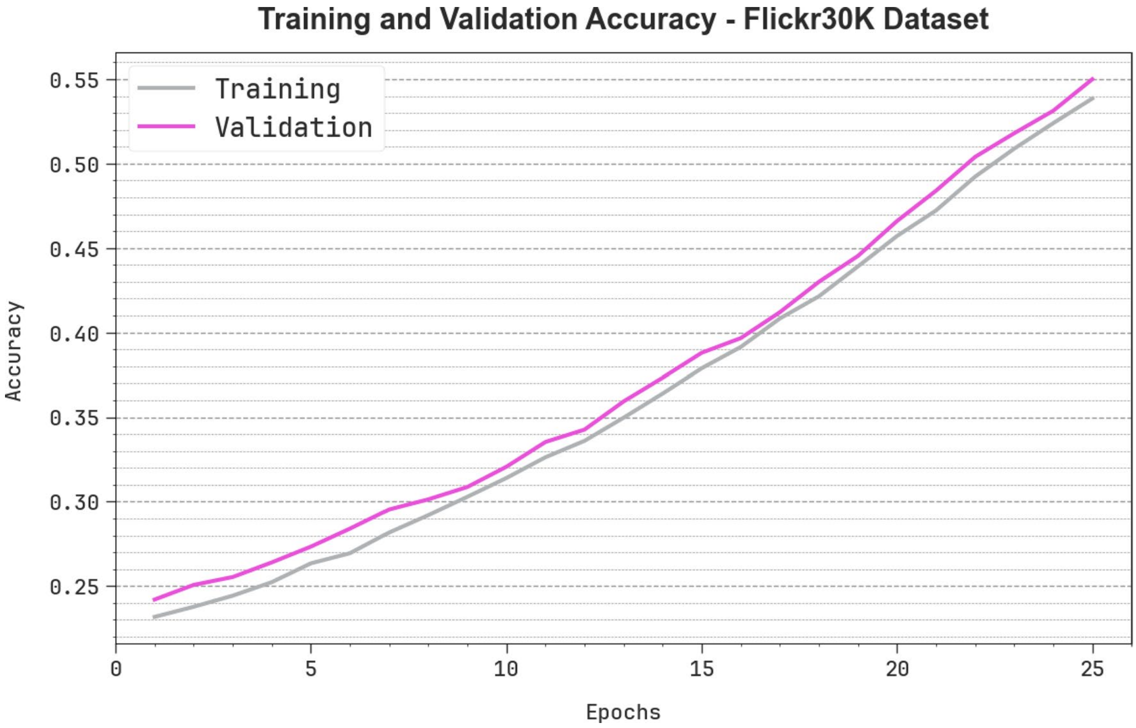

Fig. 9. $Accu_y$ curve of MARNN-FRFICP method on Flickr30K dataset.

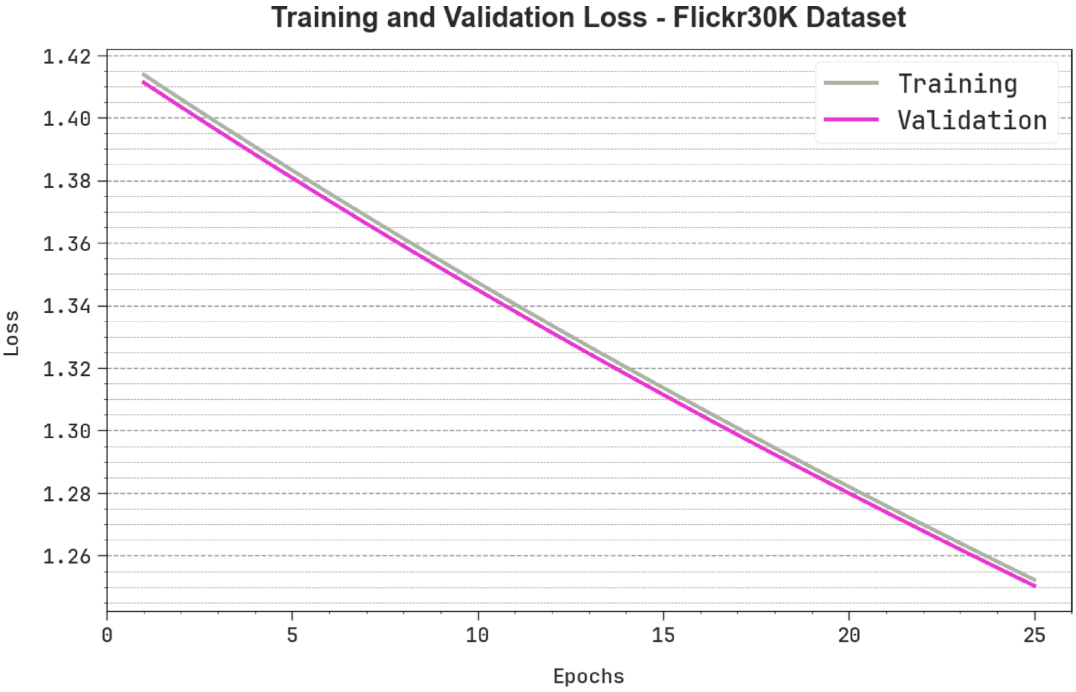

Fig. 10. Loss curve of the MARNN-FRFICP method on the Flickr30K dataset.

approach AIC-SSAIDL has gained a nearer solution in METEOR of 34.32%, and CIDEr of 138.03%. Nevertheless, existing models, such as QPULM, YOLOv8, ResNet-50, Google NIC, Soft-Attention, m-RNN, SCA-CNN-VGG, GCN-LSTM, Injection-Tag, and AIC-SSAIDL, have achieved minimal values.

In Fig. 14, the TRA $accu_y$ and VAD $accu_y$ performances of the MARNN-FRFICP technique on the MSCOCO dataset are represented. The figure highlights that both $accu_{racy}$ values exhibit a growing propensity, indicating the capability of the MARNN-FRFICP technique to achieve enhanced outcomes through repeated

Flickr30K Dataset	
Technique	**CT (sec)**
QPULM	19.72
YOLOv8	23.70
ResNet-50	12.19
Google NIC	17.19
Soft-Attention	16.38
m-RNN	24.71
SCA-CNN-VGG	15.32
GCN-LSTM	18.01
Injection-Tag	9.43
AIC-SSAIDL	14.07
MARNN-FRFICP	6.38

Table 6. CT analysis of the MARNN-FRFICP approach on Flickr30K dataset with existing models.

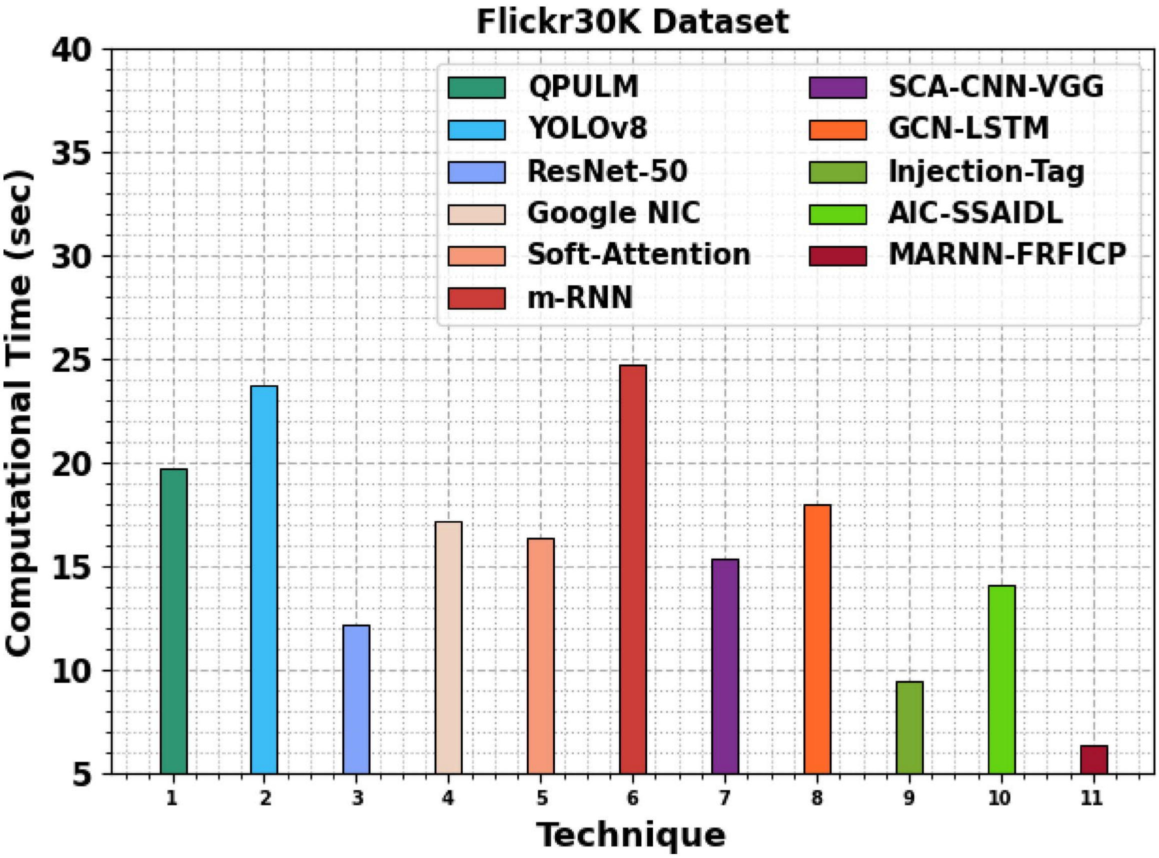

Fig. 11. CT analysis of the MARNN-FRFICP approach on Flickr30K dataset with existing models.

applications. In addition, both $accu_y$ remain relatively stable throughout the epochs, indicating decreased overfitting and demonstrating superior outcomes for the MARNN-FRFICP model, which ensures reliable predictions on unseen samples.

In Fig. 15, the TRALS and VADLS graph of the MARNN-FRFICP technique on the MSCOCO dataset is illustrated. It is exemplified that both values signify a diminishing propensity, indicating the proficiency of the MARNN-FRFICP approach in corresponding to equilibrium between generalization and data fitting. The progressive dilution in values of loss, as well as securities, secures the increased outcome of the MARNN-FRFICP approach and tunes the prediction results over time.

The CT analysis of the MARNN-FRFICP model on the MSCOCO dataset is depicted in the Table 9; Fig. 16. The MARNN-FRFICP model achieved the lowest CT of 5.71 s, outperforming other approaches such as AIC-SSAIDL at 8.64 s, YOLOv8 at 11.13 s, and Google NIC at 11.82 s. This reduction in CT indicates the high

Flickr30K Dataset						
Technique	**BLEU_1**	**BLEU_2**	**BLEU_3**	**BLEU_4**	**METEOR**	**CIDEr**
GF	72.92	65.79	65.08	54.96	41.18	65.45
InceptionResNetV2	73.47	66.63	65.59	55.60	41.85	66.06
CvT	74.33	67.34	66.15	56.19	42.65	66.85
DenseNet169	74.99	67.94	66.89	56.71	43.26	67.67
LOA	75.72	68.48	67.44	57.37	43.83	68.30
MH-BLG	76.45	69.31	68.18	58.21	44.46	69.06
MARNN-FRFICP	77.23	70.11	69.08	58.91	45.26	69.81

Table 7. Ablation evaluation of the MARNN-FRFICP technique on Flickr30K dataset.

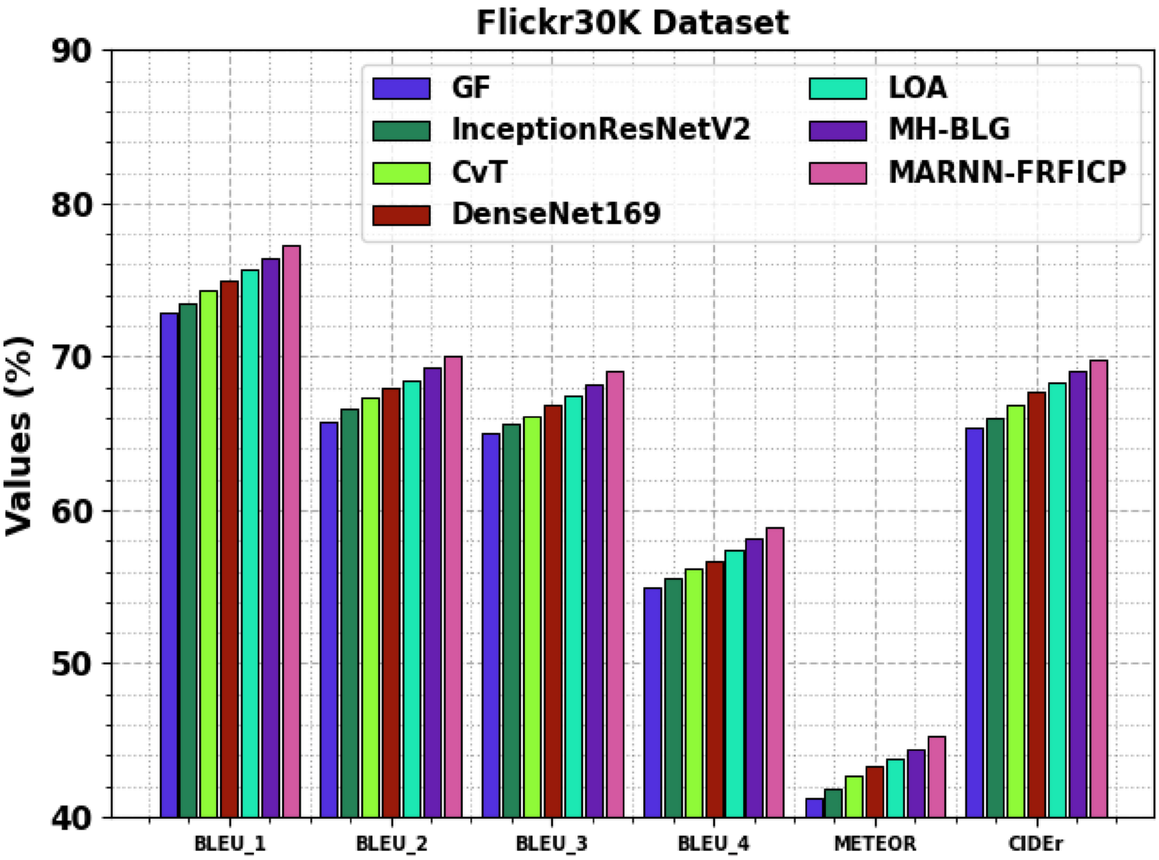

Fig. 12. Ablation evaluation of the MARNN-FRFICP technique on Flickr30K dataset.

computational efficiency and suitability of the model for large-scale image captioning tasks where speed and responsiveness are critical.

Table 10; Fig. 17 specify the ablation study of the MARNN-FRFICP model on the MSCOCO dataset. The MARNN-FRFICP model attained the highest performance across all key evaluation metrics, achieving a BLEU-4 score of 47.86, METEOR of 41.69, and a CIDEr score of 150.62. These outputs outperform prior best values like MH-BLG and LOA, indicating improved fluency, relevance, and semantic richness in the generated captions. The consistent improvements across metrics also validate the efficiency of the model's architectural components, further supported by ablation studies.

The MARNN-FRFICP approach achieved the lowest floating-point operations at 9.54 gigaflops and the least GPU memory usage at 913 megabytes, as shown in Table 11[39]. Compared to existing models such as Swin Tiny with 17.04 gigaflops and 2748 megabytes or MobileNetv3 Small with 19.57 gigaflops and 2463 megabytes, the MARNN-FRFICP model illustrates superior resource efficiency. This highlights its suitability for deployment in real time or low power environments where memory and processing constraints are critical.

MSCOCO Dataset						
Technique	**BLEU1**	**BLEU2**	**BLEU3**	**BLEU4**	**METEOR**	**CIDEr**
QPULM	63.18	49.15	34.22	23.85	20.00	69.56
YOLOv8	65.34	51.51	36.22	26.12	22.15	71.95
ResNet-50	67.55	52.92	38.17	29.08	24.87	89.45
Google NIC	63.12	49.07	34.14	23.80	19.96	69.52
Soft-Attention	65.27	51.46	36.16	26.03	22.09	71.89
m-RNN	67.49	52.85	38.09	28.99	24.81	89.37
SCA-CNN-VGG	68.21	53.61	38.86	29.65	25.34	90.06
GCN-LSTM	69.95	56.27	40.82	32.09	26.88	106.84
Injection-Tag	76.73	58.98	43.29	33.21	30.07	118.12
AIC-SSAIDL	81.14	63.65	48.59	38.83	34.32	138.03
MARNN-FRFICP	83.36	70.85	55.59	47.86	41.69	150.62

Table 8. Comparative study of the MARNN-FRFICP model with existing techniques on the MSCOCO dataset.

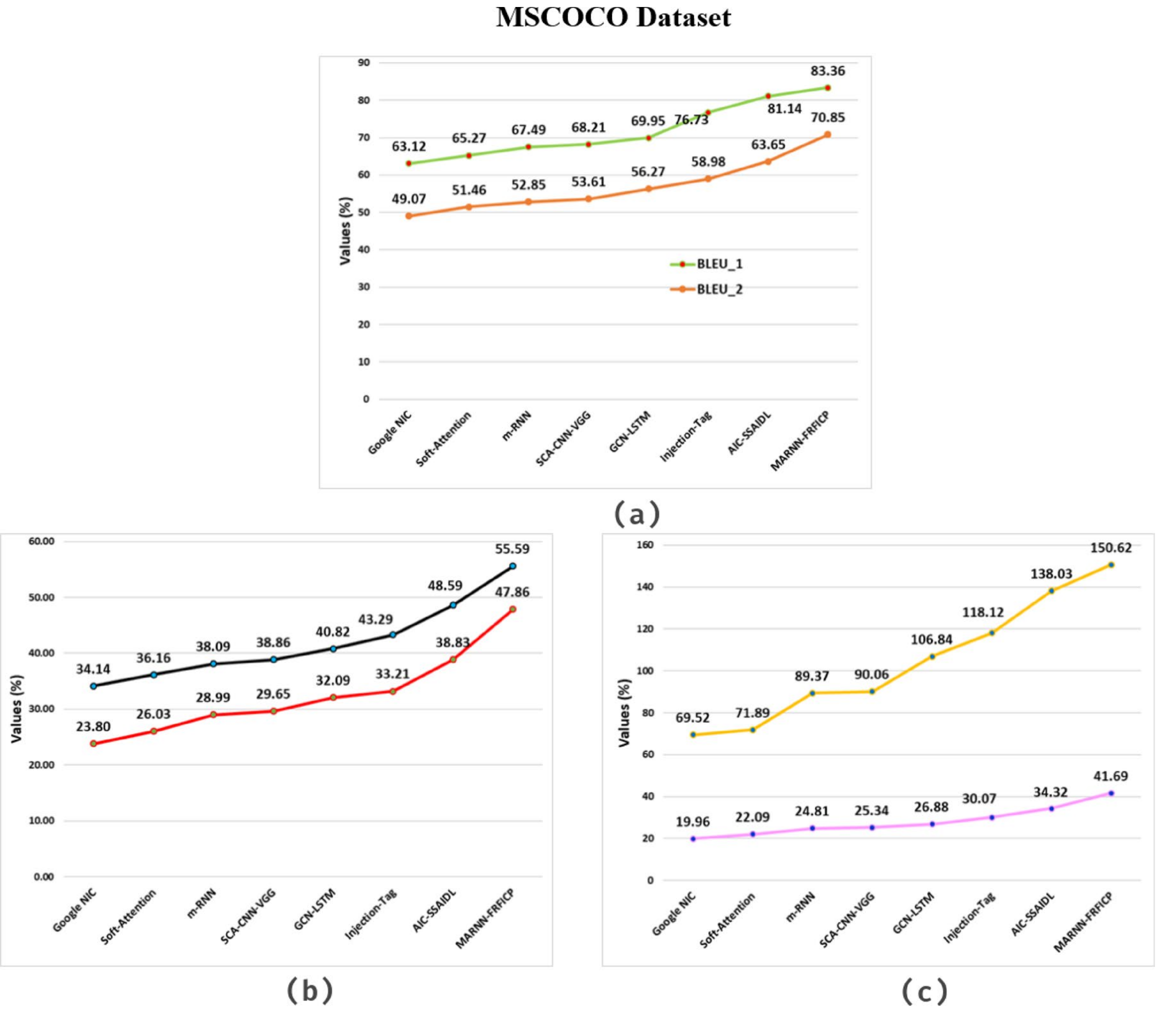

Fig. 13. Comparative analysis of the MARNN-FRFICP model on the MSCOCO dataset under various metrics.

Conclusion

In this paper, a novel MARNN-FRFICP model is proposed to assist individuals with visual impairments. The MARNN-FRFICP model aimed to enhance image captioning through an effective approach focused on improving accessibility for individuals with visual impairments, involving GF-based image pre-processing, fusion of advanced three DL models-based feature extraction, a hybrid of the MH-BLG method-based

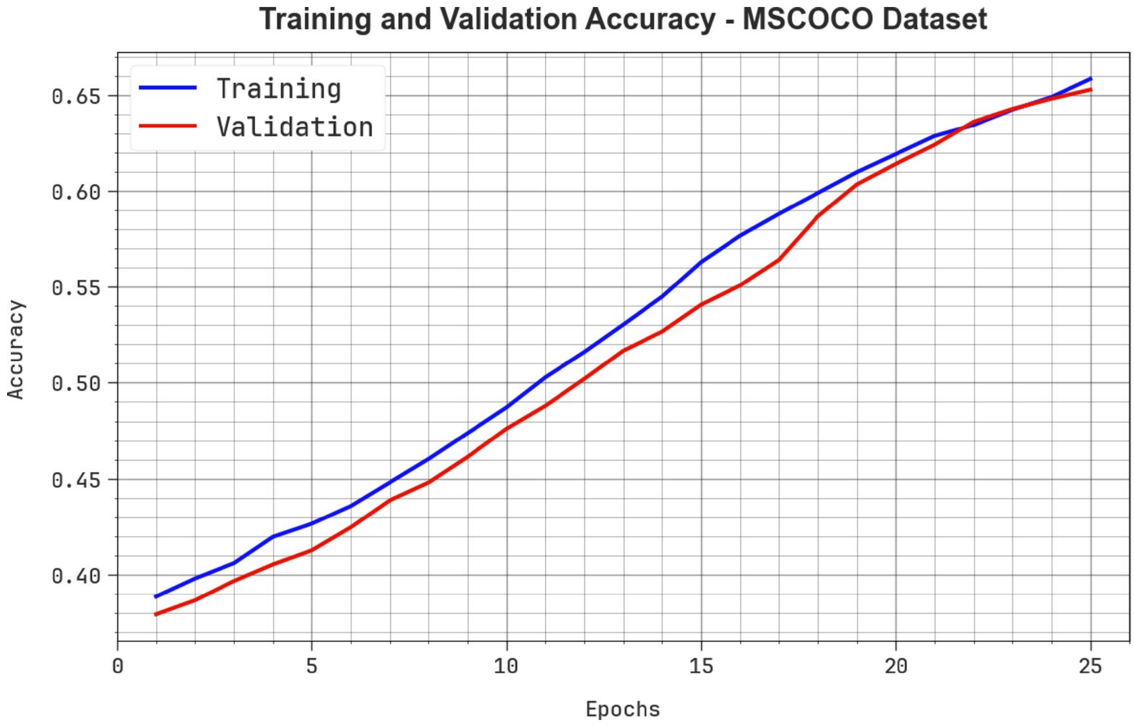

Fig. 14. $Accu_y$ curve of MARNN-FRFICP technique on MSCOCO dataset.

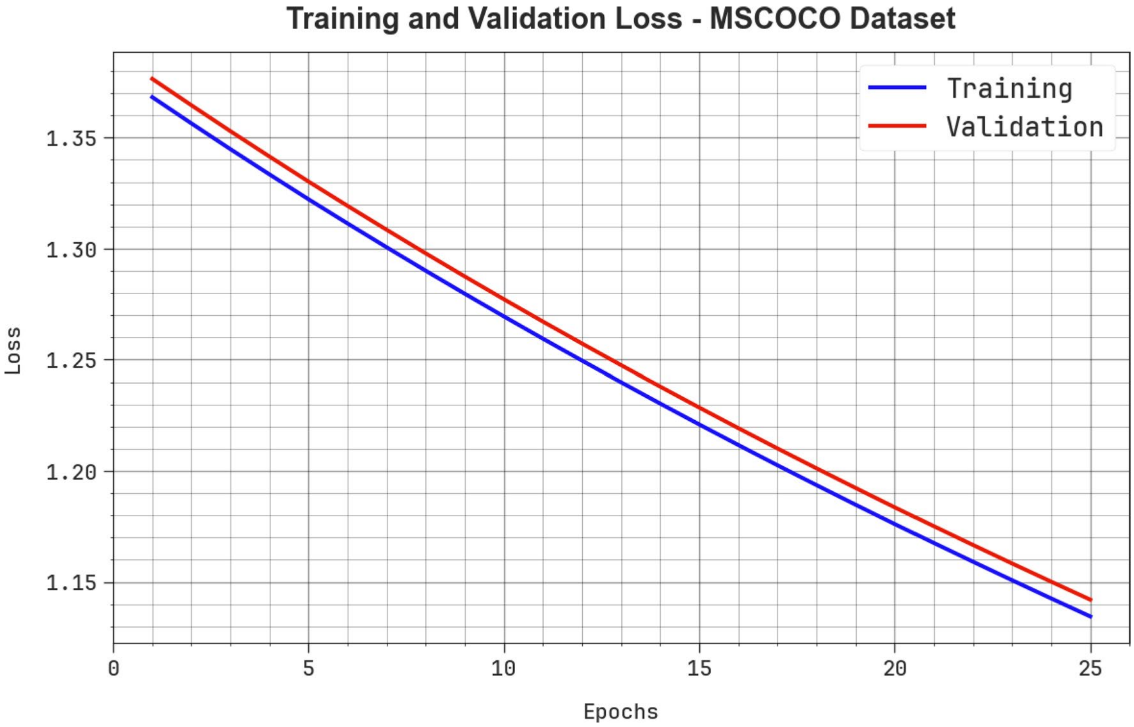

Fig. 15. Loss curve of the MARNN-FRFICP technique on the MSCOCO dataset.

classification, and LOA-based tuning. The efficiency of the MARNN-FRFICP methodology is examined under the Flickr8k, Flickr30k, and MSCOCO datasets. The experimental analysis demonstrates that the MARNN-FRFICP methodology has improved scalability and performance compared to recent techniques in various measures. The limitations of the MARNN-FRFICP methodology comprise the lack of evaluation across diverse and unseen datasets. The model does not sufficiently analyze the robustness of the technique under adversarial

MSCOCO Dataset	
Technique	**CT (sec)**
QPULM	19.45
YOLOv8	11.13
ResNet-50	19.44
Google NIC	11.82
Soft-Attention	21.95
m-RNN	19.07
SCA-CNN-VGG	13.27
GCN-LSTM	23.92
Injection-Tag	17.29
AIC-SSAIDL	8.64
MARNN-FRFICP	5.71

Table 9. CT evaluation of the MARNN-FRFICP method on MSCOCO dataset with existing approaches.

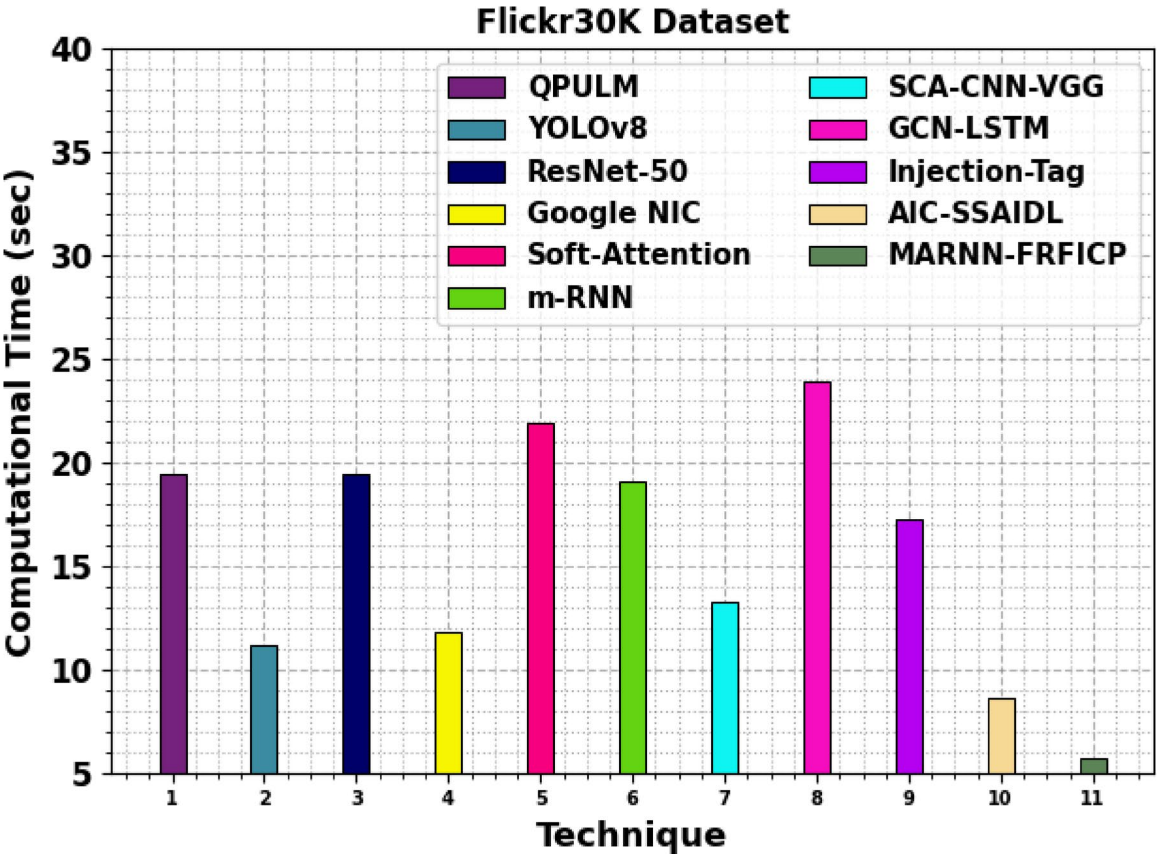

Fig. 16. CT evaluation of the MARNN-FRFICP method on MSCOCO dataset with existing approaches.

noise or varying image acquisition conditions, thus posing risks in sensitive healthcare applications. The model's generalization capability is also limited in real-world clinical settings. Moreover, while accuracy is emphasized, the explainability of predictions remains limited, which may affect trust and acceptance among clinicians. Ethical concerns such as data privacy, bias in medical datasets, and decision transparency are not thoroughly addressed. Future studies may explore domain adaptation, adversarial defence, and interpretable frameworks for broader clinical reliability.

MSCOCO Dataset						
Technique	**BLEU_1**	**BLEU_2**	**BLEU_3**	**BLEU_4**	**METEOR**	**CIDEr**
GF	79.01	67.24	51.47	43.53	37.47	146.13
InceptionResNetV2	79.76	68.02	52.06	44.34	38.22	146.96
CvT	80.56	68.56	52.66	45.13	38.97	147.57
DenseNet169	81.42	69.15	53.22	45.84	39.60	148.31
LOA	82.15	69.73	54.12	46.36	40.26	149.20
MH-BLG	82.74	70.28	54.84	47.13	41.01	149.80
MARNN-FRFICP	83.36	70.85	55.59	47.86	41.69	150.62

Table 10. Result analysis of the ablation study of the MARNN-FRFICP model on MSCOCO dataset.

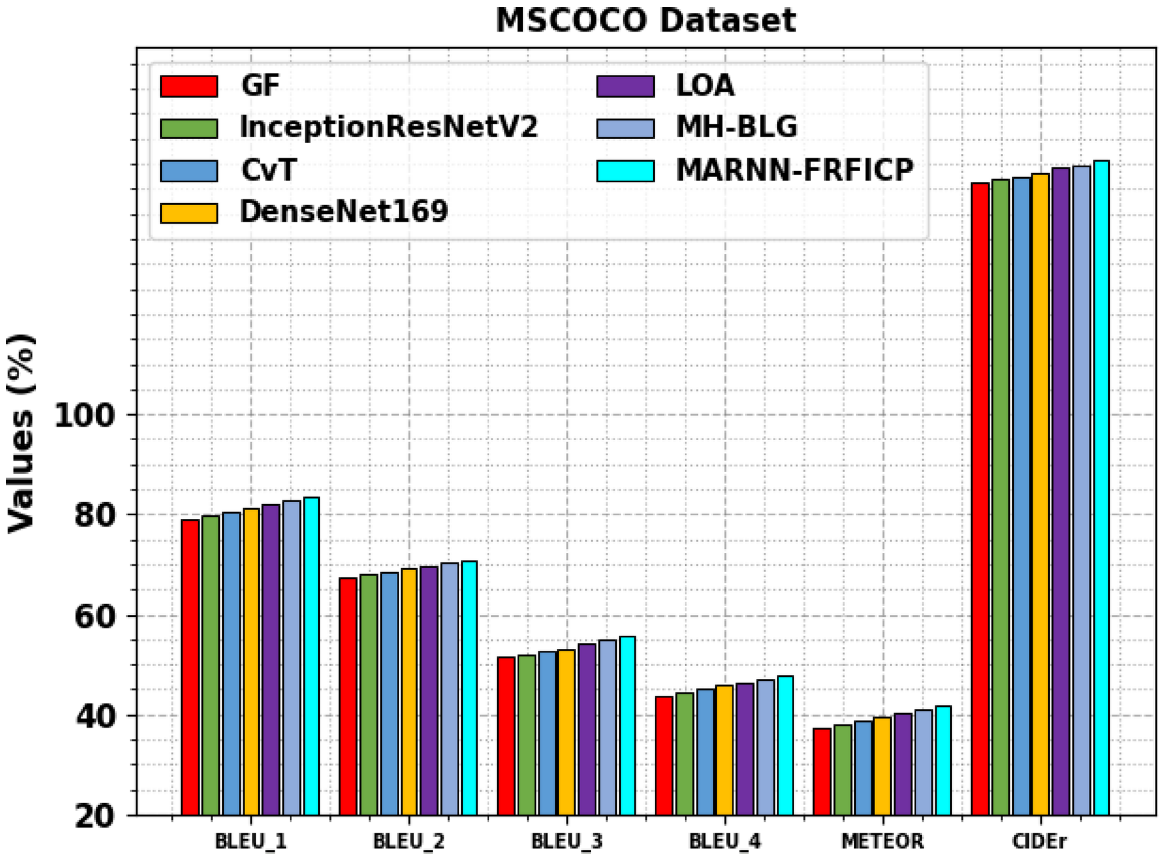

Fig. 17. Result analysis of the ablation study of the MARNN-FRFICP model on MSCOCO dataset.

Model	FLOPs (G)	GPU (M)
ConvNeXt v2 Base	26.22	1513
DenseNet 121	18.7	2189
ResNetv2 50	24.65	1660
Swin Tiny	17.04	2748
Vit Base	25.48	2710
MobileNetv3 S 50	19.57	2463
MARNN-FRFICP	9.54	913

Table 11. Comparison of model complexity of the MARNN-FRFICP technique based on flops and GPU memory usage.

Data availability

The data that support the findings of this study are openly available in the Kaggle repository at [https://www.kaggle.com/datasets/adityajn105/flickr8k](https:/www.kaggle.com/datasets/adityajn105/flickr8k) and [https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset](https:/www.kaggle.com/datasets/hsankesara/flickr-image-dataset), reference numbers[35,36].

Received: 14 June 2025; Accepted: 10 September 2025
Published online: 14 October 2025

References

1. Prashar, D., Chakraborty, G. & Jha, S. Energy efficient laser based embedded system for blind turn traffic control. *J. Cybersecur. Inform. Manage.* **2** (2), 35–43 (2020).
2. Ghandi, T., Pourreza, H. & Mahyar, H. Deep learning approaches on image captioning: A review. *ACM Comput. Surveys*. **56** (3), 1–39 (2023).
3. Sharma, H., Agrahari, M., Singh, S. K., Firoj, M. & Mishra, R. K. February. Image captioning: a comprehensive survey. In *2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)* (pp. 325–328). IEEE. (2020).
4. Feng, Y., Ma, L., Liu, W. & Luo, J. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4125–4134). (2019).
5. Herdade, S., Kappeler, A., Boakye, K. & Soares, J. Image captioning: transforming objects into words. *Advances Neural Inform. Process. Systems*, **32**. (2019).
6. Shi, Z., Zhou, X., Qiu, X. & Zhu, X. Improving image captioning with better use of captions. *arXiv preprint arXiv:2006.11807*. (2020).
7. He, S. et al. Image captioning through image transformer. In *Proceedings of the Asian conference on computer vision*. (2020).
8. Stefanini, M. et al. From show to tell: A survey on deep learning-based image captioning. *IEEE Trans. Pattern Anal. Mach. Intell.* **45** (1), 539–559 (2022).
9. Sidorov, O., Hu, R., Rohrbach, M. & Singh, A. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16* (pp. 742–758). Springer International Publishing. (2020).
10. Jain, H., Dixit, A. & Sharma, A. Detecting image spam on social media platforms using deep learning techniques. *Journal Cybersecur. & Inform. Management*, **15**(1). (2025).
11. Kalantari, F., Faez, K., Amindavar, H. & Nazari, S. Improved image reconstruction from brain activity through automatic image captioning. Scientific Reports, 15(1), p.4907. (2025).
12. Cao, Z., Xia, J. & Zhou, M. *De-confounding Feature Fusion Transformer Network for Image Captioning in Assistive Navigation Applications for the Visually Impaired* (IEEE Transactions on Instrumentation and Measurement, 2025).
13. Deepak, G. et al. K.V. and Automatic image captioning system using a deep learning approach. *Soft Computing*, pp.1–9. (2023).
14. Al Badarneh, I., Hammo, B. H. & Al-Kadi, O. An ensemble model with an attention-based mechanism for image captioning. Computers and Electrical Engineering, 123, p.110077. (2025).
15. Padate, R. & Kalla, M. *Automated Image Captioning System with Deep Learning Enabled Optimized Approach*pp.1–21 (Multimedia Tools and Applications, 2024).
16. Lee, H., Song, J., Hwang, K. & Kang, M. *Auto-Scenario Generator for Autonomous Vehicle Safety: Multimodal Attention-based Image Captioning Model Using Digital Twin Data* (IEEE Access, 2024).
17. Arasi, M. A. et al. Automated image captioning using sparrow search algorithm with improved deep learning model. *IEEE Access*. **11**, 104633–104642 (2023).
18. Deore, S. P., Bagwan, T. S., Bhukan, P. S., Rajpal, H. T. & Gade, S. B. Enhancing image captioning and auto-tagging through a FCLN with faster R-CNN integration. *Inf. Dyn. Appl.* **3**, 12–20 (2023).
19. Safiya, K. M. & Pandian, R. A real-time image captioning framework using computer vision to help the visually impaired. *Multimedia Tools Appl.* **83** (20), 59413–59438 (2024).
20. Hossain, M. I. A., Anjom, J. & Chowdhury, R. I. Towards walkable footpath detection for the visually impaired on Bangladeshi roads with smartphones using deep edge intelligence. *Array*, *26*, p.100388. (2025).
21. More, S. S. et al. Empowering the visually impaired: YOLOv8-based object detection in android applications. *Procedia Comput. Sci.* **252**, 457–469 (2025).
22. Anwar, M. et al. E-GlauNet: A CNN-Based ensemble deep learning model for glaucoma detection and staging using retinal fundus images. *Computers Mater. & Continua*, **84**(2). (2025).
23. Nguyen, H., Huynh, T., Tran, N. & Nguyen, T. MyUEVision: an application generating image caption for assisting visually impaired people. *J. Enabling Technol.* **18** (4), 248–264 (2024).
24. Qazi, N., Dewaji, I. & Khan, N. June. Vision Transformer Based Image Captioning for the Visually Impaired. In *14th International Conference on Human Interaction and Emerging Technologies: Artificial Intelligence & Future Applications, IHIET-FS 2025, June 10–12, 2025, University of East London, London, United Kingdom.* (Vol. 196, pp. 153–162). AHFE International. (2025).
25. Muhammed Kunju, A. K., Baskar, S., Zafar, S., AR, B., A, S. K. & S, R. and A transformer based real-time photo captioning framework for visually impaired people with visual attention. *Multimedia Tools Appl.* **83** (41), 88859–88878 (2024).
26. Uikey, J. et al. January. Visual Understanding and Navigation for the Visually Impaired Using Image Captioning. In *2025 International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI)* (pp. 531–536). IEEE. (2025).
27. Yousif, A. J. & Al-Jammas, M. H. A lightweight visual Understanding system for enhanced assistance to the visually impaired using an embedded platform. *Diyala J. Eng. Sciences*, pp.146–162. (2024).
28. Jenisha, J. & Priyadharsini, C. March. Enhanced Vision: Hybrid Deep Learning Based Model for Image Captioning and Audio Synthesis. In *2024 10th International Conference on Advanced Computing and Communication* Systems *(ICACCS)* (Vol. 1, pp. 2502–2507). IEEE. (2024).
29. Biswas, A. & Branicki, M. A unified framework for the analysis of accuracy and stability of a class of approximate Gaussian filters for the Navier–Stokes equations. *Nonlinearity*, *37*(12), p.125013. (2024).
30. Rastogi, D. et al. Brain Tumor Detection and Prediction in MRI Images Utilizing a Fine-Tuned Transfer Learning Model Integrated Within Deep Learning Frameworks. *Life*, *15*(3), p.327. (2025).
31. Putra, B. P. E., Prasetyo, H. & Suryani, E. Residual Transformer Fusion Network for Salt and Pepper Image Denoising. *arXiv preprint arXiv:2502.09000*. (2025).
32. Hemal, M. M. & Saha, S. Explainable deep learning-based meta-classifier approach for multi-label classification of retinal diseases. *Array*, p.100402. (2025).
33. Yu, Y., Yang, H., Peng, F. & Wang, X. Drilling Condition Identification Method for Imbalanced Datasets. *Applied Sciences*, *15*(6), p.3362. (2025).

34. Xi, H. et al. Prediction of Lithium Battery Voltage and State of Charge Using Multi-Head Attention BiLSTM Neural Network. *Applied Sciences*, *15*(6), p.3011. (2025).
35. https://www.kaggle.com/datasets/adityajn105/flickr8k
36. https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset
37. Lin, T. Y. et al. September. Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740–755). Springer, Cham. (2014).
38. Bae, J. W., Lee, S. H., Kim, W. Y., Seong, J. H. & Seo, D. H. Image captioning model using part-of-speech guidance module for description with diverse vocabulary. *IEEE Access.* **10**, 45219–45229 (2022).
39. Redondo, A. et al. Multiclass classification of oral mucosal lesions by deep learning from clinical images without performing any restrictions. *Biomedical Signal Processing and Control*, *111*, p.108337. (2026).

Acknowledgements

The authors extend their appreciation to the King Salman center For Disability Research for funding this work through Research Group no KSRG-2024-083.

Author contributions

Mashael M Asiri: Conceptualization, methodology, validation, investigation, writing—original draft preparation, Kholoud Alghamdi: Conceptualization, methodology, writing—original draft preparation, writing—review and editing. Fahad Alzahrani: methodology, validation, writing—original draft preparation. Mahir Mohammed Sharif: software, visualization, validation, data curation, writing—review and editing.

Funding

The authors extend their appreciation to the King Salman Centre for Disability Research for funding this work through Research Group no. KSRG-2024- 083.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.M.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.