International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I3P112

Volume 74 Issue 3, 153-168, March 2026
© 2026 Seventh Sense Research Group®

Original Article

ViT-NARCap: Image Captioning with Vision Transformer Context-Aware Nucleus Sampling and Roberta Re-Ranker

Bhargavi Polepalli[1], Praveen Kumar Sekharamantry[2], Konda Srinivasa Rao[3]

[1,2,3]*GST, GITAM (Deemed to be University), Department of CSE, Visakhapatnam, India.*

[1]*Corresponding Author : bpolepal@gitam.in*

Received: 01 July 2025 Revised: 13 October 2025 Accepted: 17 November 2025 Published: 28 March 2026

***Abstract** - Image captioning is a semantically correct and linguistically competent attempt to produce textual captioning based on visual art, but it is difficult as it is restricted by contextual knowledge and language variety. The traditional encoder-decoder systems that are usually founded on CNN encoders and RNN decoders have problems with long-range dependency estimation, exposure bias, and repetitive captioning. In response to these shortcomings, this paper suggests a superior image captioning network that combines a Vision Transformer encoder and a normalized auto-regressive fine-tuned Transformer decoder. The Vision Transformer is one of the better techniques for capturing hierarchical visual representations and global spatial relationships. In contrast, the transformer-based decoder holds together coherent and context-aware sentence generation. To increase further the diversity and fluency of captions, the idea of nucleus sampling is used in the process of decoding, and a reranking mechanism based on the use of RoBERTa is presented to fine-tune the selected captions in accordance with semantic relevance. The results of the experimental Analysis of the benchmark datasets indicate that the offered approach tends to be superior to the current ones in standard measures, such as BLEU, CIDEr, ROUGE-L, and METEOR, which proves its efficiency and strength.*

***Keywords** - Recurrent Neural Network (RNN), Convolutional Neural Networks (CNN), Image Captioning, Nucleus Sampling, and Vision Transformer.*

1. Introduction

The era of Artificial Intelligence over the last few years has transformed the technology of computer vision and natural language processing significantly, and machines are now able to gain visual data and generate textual descriptions that are similar to those that humans give [1]. The Image captioning, the problem of providing a semantically rich and grammatically correct description of an image, has been a significant vision-language problem that has been applied in the accessibility system, multimedia search, assistive technology to aid blind users, human-computer interaction, and intelligent surveillance. Compared to image classification, which gives a picture only one label, image captioning demands a deeper perception of visual scenes, interrelations between objects and acts, and contextual information, and is a much more difficult and cognitively challenging question [2, 3]. Captioning methods used in the first image were rule-based and visual features that were hand-drawn and could not be generalized, nor did they extract semantic richness [4]. Everything also changed with the time of Deep Learning, particularly encoder-decoder models, which are a combination of Convolutional Neural Networks to extract visual features and recurrent Neural Networks to produce sentences [5]. Even though these models present attractive performance, they have limitations that are inherent to them, such as a lack of modeling long-term dependency, exposure bias in training, and the generation of generic or repetitive captions. Also, recurrent architectures implement sequences sequentially, and consequently inhibit parallelism and, in most instances, inefficient acquisition of contextual relationships at a global scale [6].

In order to minimize these limitations, attention mechanisms were introduced, which enable the models to attend to salient regions of images during captioning. In spite of the fact that attention-based CNN-RNN models might be more effective in aligning visual features with the generated words, they were confined by convolutional encoders due to their inability to learn global spatial correlations. Most recently, transformer-based design has gained popularity due to its capability to self-attend, allowing parallel processing and improved management of long-range dependencies. Vision Transformers have particularly demonstrated a high capability in accessing global contextual information by treating images as a sequence of patches, contrary to localized convolutional features [7]. There are several outstanding challenges with image captioning methods based on transformers with these developments. To start with, most strategies have many, and they concentrate on the originality of the architecture, and they do not engage much research on the linguistic diversity. In most instances, they end up with captions that are not only fluent but also redundant.

Second, greedy decoding or beam search methods that most caption generation pipelines probe encourage the use of high-probability words and, therefore, choose to limit creativity, which yields safe and boring outputs. Third, in contrast to the works that feature trained language models, they are usually used individually to encode or decode, and there is no straightforward method of reranking and refining generated captions, depending on their relevance to the context and fluency. Finally, the recent studies are not only documenting the competitive quantitative results but failing to provide the interpretative Analysis of the reasons why these performance improvements are obtained in that way, limiting their practical applicability and the reproducible consistency [8].

Such shortcomings indicate an image captioning model that would simultaneously support image perception on a global scale, language and cultural diversity, and focus on particular refinement, and provide a comprehensible and actually well-grounded performance analysis. The joint optimization of Vision Transformer to hierarchically encode visual information, controlled probabilistic sampling to boost diversity, and large-scale pre-trained language models to refine captions has not been adequately researched in the literature to optimize captions in one architecture.

The present study is aimed at improving upon these weaknesses by introducing a new image captioning architecture, which combines a Vision Transformer encoder and a normalized auto-regressive fine-tuned Transformer decoder using context-sensitive nucleus sampling and reranking based on a RoBERTa-formed robustness model [9]. Unlike the conventional encoder-decoder models, the given architecture will acquire fine-grained and coarse-scale spatial dependencies with hierarchical self-attention, and it will be possible to construct more expressive visual representations.

In addition, high-probability tokens are discouraged by nucleus sampling, which easily supports linguistic diversity and cannot ruin the contextual relevance. Another feature that makes this work stand out from the existing solutions is the presence of a RoBERTa-based reranking component that suggests explicitly refining the decisions of candidate captioning with the help of semantic sense and language fluency rather than decoder probabilities [10]. As opposed to the earlier study, which focuses on either architectural improvements or decoding processes in isolation, the present study provides an elaborate vision-language framework that addresses the fundamental limitations of the transformer-based image captioning in a systematic way. Large-scale experimental trials on benchmark data demonstrate steady performance improvements on a set of evaluation measures, and Analysis is performed in detail to characterize the input of each of the architectural components to performance improvements. Through this, the paper leads to a state-of-the-art breakthrough in image captioning since more accurate, varied, and context-related captionings are presented, which enhances the potential of captioning systems in practice [11].

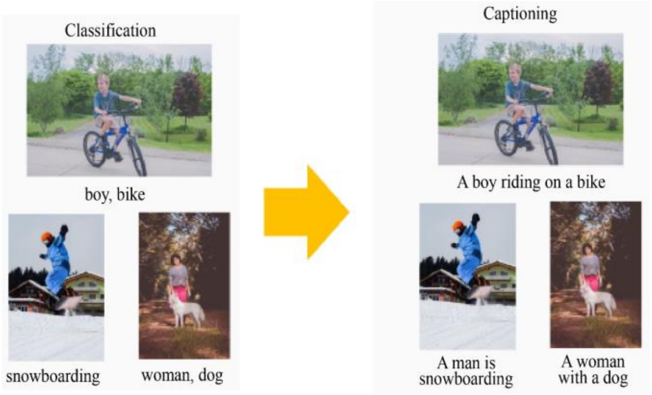

Fig. 1 Image caption and image classification

The existing works on image captioning can be categorized into CNNRNN models, attention-enhancing designs, transformer-based designs, and diversity-oriented captioning designs. The CNNRNN models were built on the foundation of mapping images to text sequences. However, because they relied on recurrent decoding, which was the primary basis of their implementation, they were less scalable and had less contextual modeling. The attention-based extensions improved the matching of object words, although they were restricted by convolutional receptive fields.

Many of these shortcomings were overcome with the help of transformer-based models since they enabled attention on the world rather than visual tokens. The CT Vision Transformer Captioning systems had improved context sensitivity and parallel processing features. However, the majority of the reported systems have deterministic decoding algorithms, which are sensitive to probability maximization rather than linguistic expressiveness. This results in captions being regarded as lacking descriptive richness in cases where there are high scores in metrics.

Also, although the image caption refining and reranking capabilities of pre-trained language models, such as BERT and RoBERTa, have not been thoroughly examined to date, they show impressive results on the tasks of natural language understanding. Recent studies have attempted to improve caption diversity through stochastic decoding. However, these methods are most often applied without semantic validation, with the consequence of generating grammatically correct but visually inappropriate captions. Furthermore, comparative Analysis tends to emphasize the numerical returns, and in doing so has not sufficiently addressed the failure incidences, part-by-part contributions, and interpretability, and has left the key questions of paramount importance unanswered. The proposed work is remarkable since it suggests addressing these limitations in one way. The framework being applied is a combination of hierarchical Vision Transformer encoding, normalized auto-regressive decoding, probabilistic nucleus sampling, and semantic reranking on a pre-trained RoBERTa model, as opposed to other techniques, which are only trained as individually working algorithms. Not only does this general design increase the quantitative performance, but also the qualitative caption quality, interpretability, and strength.

Despite the recent progress in image captioning that has helped a great deal in aligning vision and language, various architectural and decoding drawbacks are yet to be addressed. In order to put these challenges into perspective and determine the gaps in knowledge that exist on the subject, the following discussion presents a full review of the state-of-the-art image captioning methodologies. The following sections comprise the remainder of the article: The description of recent image captioning studies using deep learning-based techniques is shown in Part II. The suggested deep learning-based solution is shown in Section III. The results of the suggested methodology are presented in Section IV. Lastly, the concluding observations are presented in Section V.

2. Literature Review

The major trends of the existing picture captioning methods based on Deep Learning are briefly mentioned in this portion. As aforementioned, a significant amount of interest has been drawn towards image captioning, as the method of providing natural language descriptions to images, due to its capability of enhancing human-computer interaction, image organization, and image retrieval. The recent breakthroughs in deep learning have triggered spectacular advances in this area, with powerful structures and innovative methods to enhance cross-modal understanding, contextuality, and the quality of captions. This literature review discusses the development of image captioning models, encoder-decoder structures, attention, ViTs, GANs, and language-specific models to deal with issues in multilingual settings. The underlying image captioning model is the encoder-decoder model, which obtains image features through CNNs and provides captioning through RNNs. Verma et al. [18] suggested an encoder-decoder model with VGG16 Hybrid Places 1365 as an encoder and LSTM as a decoder. It was trained on datasets of Flickr8k and MS-COCO Captions, and it had high performance with a BLEU-1 score of 0.6666 on Flickr8k and 0.7350 on MS-COCO. Their work underlines the ability of CNN-RNN structures to generate grammatically correct captions. In the same way, Alzubi et al. [16] introduced an ensemble system incorporating the use of Inception and a 2-layer LSTM, with other experiments involving the Gated Recurrent Unit (GRU) and Bidirectional LSTM variants. The model was used to improve caption generation, leveraging GloVe embeddings of word vectors on the Flickr8k dataset with the highest BLEU-4 score of 55.8%. This paper will use CNNs to extract features and sequential models to generate captions.

Attention mechanisms have been incorporated in image captioning models to enhance the contextual background and object recognition. The Bengali image captioning task was presented by Bhuiyan et al. [19] using a context-sensitive attention mechanism with a bidirectional GRU to provide practical solutions to the issues of object-word association and context preservation. Their model demonstrated outstanding gain in the METEOR scores on three of the larger benchmark datasets and demonstrates the significance of the attention mechanisms in multilingual environments. Mishra et al. [20] have implemented an encoder-decoder model that has an efficient channel attention mechanism to use in Hindi image captioning. Using the ECA-NET CNN and the Bahdanau attention, the model focused on image channels selectively, which improved the process of feature extraction and contextual relevance. This method brought significant changes in BLEU scores, and it shows the effectiveness of channel attention mechanisms in language-specific conditions.

The current developments in ViTs and GANs have driven image captioning even further. Tyagi et al. [21] presented the ICTGAN model as a combination of ViTs and GANs to learn representations, as well as generate realistic captions. This model was used to combine textual and visual features by a self-attention mechanism, which gave better relevance, diversity, and coherence of captions on the MS-COCO and Flickr30k data sets. The strong generalization abilities of this method indicate the potential of ViTs and GANs in complicated image-caption settings. In order to solve the difficulty of caption generation in non-English languages, some models have been built based on customized architectures.

The fuzzy attention-based DenseNet-BiLSTM framework for image captioning in Chinese was introduced by Lu et al. [22]. This strategy has improved the global and detailed feature extraction with DenseNet and a fuzzy attention mechanism that has solved the problem of contextual alignment. This model was tested on the AI Challenger dataset, where it was found to be more effective at various metrics, which is why it is effective in dealing with complex semantic structures. In the same light, Bhuiyan et al. [19] and Mishra et al. [20] addressed the issues of Bengali and Hindi caption generation, respectively, with the help of context-sensitive attention and effective channel attention models. These works underscore the need for specialized architectures and attention mechanisms for low-resource languages.

Table 1. Comparative review of existing image captioning approaches

Ref.	Year	Visual Encoder	Language Decoder	Key Technique	Dataset(s)	Strengths	Identified Limitations
Verma et al.	2024	CNN (VGG16)	LSTM	Encoder–Decoder	Flickr8k, MS-COCO	Grammatically correct captions, stable training	Weak global context modeling, repetitive captions
Alzubi et al.	2021	CNN (Inception)	LSTM / GRU	Ensemble Learning	Flickr8k	Improved robustness via ensemble	High computational cost, limited diversity
Bhuiyan et al.	2024	CNN	Bi-GRU	Context-aware Attention	Bengali datasets	Better object–word alignment	Language-specific, limited scalability
Mishra et al.	2021	CNN (ECA-Net)	LSTM	Channel Attention	Hindi datasets	Enhanced feature relevance	Limited global spatial understanding
Lu et al.	2021	DenseNet	Bi-LSTM	Fuzzy Attention	AI Challenger	Improved semantic alignment	Sequential decoding inefficiency
Tyagi et al.	2024	Vision Transformer	GAN-based	ViT + GAN	MS-COCO, Flickr30k	Diverse and realistic captions	Training instability, high complexity
Elbedwehy et al.	2022	Vision Transformer	Transformer	Full Transformer	MS-COCO	Strong contextual modeling	Greedy decoding causes repetitive captions
Katpally et al.	2020	CNN	LSTM	Ensemble Networks	Flickr8k	Improved metric scores	Lack of interpretability
Ma et al.	2023	CNN + Local Attention	Transformer	Local Visual Modeling	MS-COCO	Better region-level descriptions	Limited linguistic diversity
Proposed Model	2025	Vision Transformer (Hierarchical)	Normalized Auto-Regressive Transformer	Nucleus Sampling + RoBERTa Reranking	Flickr8k, Flickr30k	High contextual accuracy, diverse captions, and semantic refinement	Moderate computational overhead

The recent research has made significant progress in vision-language modeling and image captioning to enhance the efficiency of transformers, their flexibility, and their ability to understand semantics. Su et al. proposed RoFormer that uses rotary position embeddings to improve the positional encoding of transformers to improve modeling of long-range dependencies at minimal computational cost [40]. Wang et al. introduced LoRA-GA, a low-rank adaptation approach, which uses the strategy of gradient approximation, to achieve competitive performance and significantly lower the training complexity in order to enhance parameter-efficient fine-tuning further [41]. Regarding remote sensing image captioning, Yang et al. introduced a multi-attentive network with diffusion models to produce more varied and more context-sensitive change captions, which have a higher semantic richness [42]. In line with this, Zhu et al. introduced Semantic-CC, which leverages prior knowledge and semantic directions to improve caption accuracy and readability in remote sensing change captioning tasks [43]. Bai et al. described Qwen2.5-VL, a large multimodal foundation, as the study of visual perception and language reasoning, as a trend toward scalable and general-purpose vision-language systems, on a broader vision-language scale [44]. Recently, Wang et al. introduced SAT-Cap, a single-stage transformer-based remote sensing change captioning network that compresses the captioning pipeline without a significant drop in performance, indicating a move towards end-to-end and efficient transformer architectures [45].

According to the Analysis of literature, current methods are inclined to treat the architectural design, decoding strategies, or diversity addition separately. Based on these observations, the following section provides a single image captioning framework, which integrates visual encoding, controlled decoding, and semantic reranking into a single architecture.

3. Proposed Model

This section presents the detailed discussion about the proposed approach, where the first subsection presents the overview of the proposed architecture, subsection II presents the discussion about the data preprocessing phase, where image scaling, normalization, and image augmentation tasks are performed on the image data, and where the textual preprocessing is also performed on the raw caption data. Next, subsection III presents the discussion on the InceptionV3 DL architecture for feature extraction. Later, the obtained features are then processed through the encoder module, where we introduced a Modified Hierarchy Multi-Context Attention Vision Transformer encoder module. Later, this encoded data is processed through the Normalized Auto-Regressive Fine-Tune Transformer Layer Decoder module. Finally, the RoBERTa-based Reranking Model is applied to generate the captions, and the nucleus sampling mechanism is also incorporated to enhance the diversity.

3.1. Overview of Proposed Architecture

The suggested model combines the idea of deep learning and transformer-based models to produce captions of high quality in relation to image data. It is based on a multi-stage pipeline that allows successful feature extraction, encoding, and decoding with the use of attention mechanisms and language modeling to enhance caption generation. This entire process is divided into several steps, and they are discussed below:

3.1.1. Data Preprocessing

Preprocessing of data is performed in order to enhance the quality and durability of both text and visual data before the model training. Image Scaling & Normalization: Image varies in resolution, so by rescaling and normalizing them, there is consistency, and it makes the model efficient in processing the inputs. Normalization ensures a quicker model convergence when training a model because it standardizes pixel values and leads to improved generalization.

Image Augmentation: Image Augmentation methods that include rotation, flipping, cropping, and color jittering are employed to make data more diverse and to minimize overfitting and enhance model robustness, especially where there is a limited supply of labeled data.

Textual Preprocessing: Raw caption data is text-normalized, lower-cased, and punctuated, and then tokenized. These measures contribute to the noise of the dataset reduction and make sure that the text entries are organized in a manner that the model will be able to process them and learn.

3.1.2. Feature Extraction

The InceptionV3 deep learning structure is used to produce high-level features on images. This architecture is effective in capturing multi-scale spatial information with the inception modules, hence it is suitable for extracting rich visual features. In addition, it has also been trained on large-scale datasets, and thus it can transfer learn effectively without significant training. The features used therefore use a condensed form of the images, which makes them easier to compute yet retains significant information.

3.1.3. Encoding Module

Following the extraction of features, the extracted features are fed into a Modified Hierarchy Multi-Context Attention ViT encoder module that boosts the capability of the model to comprehend intricate spatial connections in the picture. The classic CNN models emphasize local spatial organization compared to ViTs, which are able to highlight the long-range dependence of images, and thus can be used to understand a scene better. On the same note, it employs a multi-Context Attention module that is used to enhance the capacity of the model to concentrate on important areas of an image on various levels of abstraction so as to generate a better representation of features. Also, the features derived are represented in an organized hierarchy that ensures that the small details as well as big picture information are also preserved and enhance caption accuracy.

3.1.4. Decoder Module

The proposed Normalized Auto-Regressive Fine-Tune Transformer Layer Decoder is used to decode the data with the help of the encoded attributes. This autoregressive character of the proposed decoder assists in predicting words sequentially, with the words that have been generated conditioning the generation of the next word, and this is what makes them become coherent and grammatically correct. Moreover, during training, it alters and improves the learned representations, enabling the model to generate more diverse and context-appropriate captions.

3.1.5. Caption Generation

Finally, to ensure that the generated captions are both syntactically and semantically relevant, the technique incorporates a RoBERTa-based Reranking Model on the generated caption text. It enhances language representation by leveraging large-scale pre-training and dynamic masking, making it more effective than traditional LSTMs or standard Transformers.

Also, it considers a reranking mechanism to ensure that the final output aligns closely with the image content. The traditional ranking methods always select the highest probability words, which can make captions repetitive; therefore, the technique incorporates a nucleus sampling model to select the solution from the probability distribution to ensure greater diversity and creativity in generated captions.

3.2. Data Preprocessing Module

This section presents a detailed discussion about the data preprocessing phase, where we have performed several preprocessing steps on image and caption data. This includes image resizing, color normalization, and data augmentation, whereas the text preprocessing includes tokenization, stop word removal, Lemmatization, and Vectorization (Word Embedding).

3.2.1. Image Data

The image resizing involves transforming an image I of original size $H \times W$ into a new desired size. $H' \times W'$. This operation is expressed in Equation 1 below:

$$I' = R(I, H', W') \tag{1}$$

Where, $I \in \mathbb{R}^{H \times W \times C}$ denotes the original image with height H, width W, and C channel, and I' is the resized image and $R(.)$ is the resizing function where bicubic interpolation or bilinear interpolation are commonly used. The resized image can be obtained as:

$$I'(x', y') = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} I(x, y).\, w(x, x').\, w(y, y') \tag{2}$$

Similarly, the color normalization is a process to standardize pixel values across different images, improving model convergence. For any image I', the normalized pixel values can be obtained as:

$$I_n = \frac{I' - \mu}{\sigma} \tag{3}$$

Where I_n is the normalized image, μ and σ are the mean and standard deviation of pixel values. It is performed channel-wise for RGB images, and can be expressed as

$$I_n^{(c)} = \frac{I'^{(c)} - \mu_c}{\sigma_c}, c \in \{R, G, B\} \tag{4}$$

Where μ_c and σ_c are the mean and standard deviation of the color channel c. On this data, the augmentation processes are applied, including random cropping and color jittering. During the random cropping phase, random crop extracts a sub-image I_c of size $H_c \times W_c$ from I' where the coordinates to crop the image are (x_s, y_s), expressed as:

$$I_c = C(I', x_s, y_s. H_c, W_c) \tag{5}$$

Where (x_s, y_s) represents the randomly selected coordinates and $C(.)$ This is the cropping function. Similarly, the color jittering creates random variations in brightness, contrast, saturation, and hue. The transformation is defined as:

$$I_j = \alpha I' + \beta \tag{6}$$

Where I_j represents the jittered image, $\alpha \sim \mathcal{U}(a_{min}, a_{max})$ and $\beta \sim \mathcal{U}(b_{min}, b_{max})$ adds a random shift in intensity.

3.2.2. Textual Data

As discussed before, the raw captions are in the form of textual data, which need to be preprocessed to enhance the system performance. In this work, the technique has been performed following the preprocessing steps:

Tokenization: Given a raw caption S, tokenization splits it into a sequence of words or subwords as follows:

$$S = \{w_1, w_2, \dots, w_n\} \tag{7}$$

Where w_i represents the individual tokens. In the next step, stopword filtering is performed where A set of stopwords $\mathcal{W}$ is predefined, and each word is filtered out if $w_i \in \mathcal{W}$. Finally, vectorization or word embedding is performed, where each word is mapped to a high-dimensional vector representation v_i using an embedding function f_e.

This can be expressed as:

$$v_i = f_e(w_i) \tag{8}$$

Where f_e Word2Vec embeddings. The outcome of these steps can be as follows:

Raw caption	Tokenization	Stopword removal
This is the image caption work.	S {"This", "is", "the", "image", "caption", "work"}	S′={"image", "caption", "work"}
Vectorization V={"image"→[0.42,0.85,0.13,0.92],"caption"→[0.56,0.33,0.78,0.61],"work"→[0.89,0.50,0.44,0.21]}		
$X = \begin{bmatrix} 0.42 & 0.85 & 0.13 & 0.92 \\ 0.56 & 0.33 & 0.78 & 0.61 \\ 0.89 & 0.50 & 0.44 & 0.21 \end{bmatrix}$		

3.3. Feature Extraction

This section presents the feature extraction process, where the technique uses the InceptionV3 model to extract the features from images. The Inceptionv3 model extracts hierarchical attributes using multiple convolutional pathways. These convolution modules consist of 1x1, 3x3, and 5x5 convolutions, followed by max-pooling and average-pooling operations. The initial 1x1 Convolution is used to perform the dimensionality reduction. Its operation can be expressed as:

$$F_{1\times1} = \sigma\left(W_{1\times1} * I^{"} + b\,\right) \tag{9}$$

Where $*$ represents Convolution, $W_{1\times1}$ is the weight matrix for 1x1 Convolution, b is the bias term, and σ is the activation function. Further, 3×3 and 5×5 convolutions are used for feature extraction as

$$F_{3\times3} = \sigma\left(W_{3\times3} * I^{"} + b\right) \quad (10)$$

$$F_{5\times5} = \sigma\left(W_{5\times5} * I^{"} + b\right) \quad (11)$$

Later, max-pooling and convolutional operations are performed, where max-pooling helps to reduce the Spatial Feature and retains the most important features. Later, convolutional operations are performed on the max-pooling outcome for feature compression and learning the non-linear representation. These two operations can be represented as follows:

$$F_{pool} = MaxPool\left(I^{"}\right) \quad (12)$$

$$F_{pool\ 1\times1} = \sigma\left(W_{1\times1} * F_{pool} + b\right) \quad (13)$$

Thus, Max-Pooling extracts dominant features while reducing spatial dimensions, while 1×1 Convolution reduces computational cost and refines features before further processing.

These operations help InceptionV3 efficiently learn multi-scale visual features. These convolutions and pooling operation outcomes are concatenated to obtain the feature vector as

$$F_{inception} = Concat\left(F_{1\times1}, F_{3\times3}, F_{5\times5}, F_{pool-1\times1}\right) \quad (14)$$

After passing through multiple Inception modules, the feature maps are reduced to a fixed-length vector using Global Average Pooling (GAP), which is expressed as:

$$F_{GAP}(c) = \frac{1}{H'W'}\sum_{x=1}^{H'}\sum_{y=1}^{W'} F_{inception}(x, y, c) \quad (15)$$

Here, F_{GAP} represents the feature vector v_iBased on this, the final feature vector can be represented as:

$$v = [v_1, v_2, v_3, \dots v_n] \quad (16)$$

Where n represents the output of the final pooling layer, the following figure depicts the architecture of InceptionV3 for feature extraction. The obtained features are then fed to the feature encoder and decoder module.

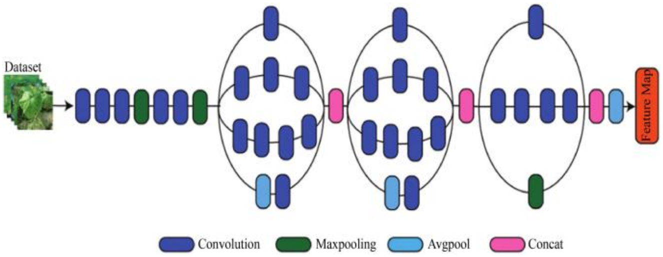

Fig. 2 Overall feature extraction pipeline using convolution, pooling, and multi-scale aggregation

3.4. Encoder and Decoder Module

This section describes the proposed encoder and decoder module to enhance the image captioning performance.

3.4.1. Encoder Module

The encoder module encrypts these characteristics using the features that were extracted from the InceptionV3 model. To capture the long-range dependencies, the technique has created a hierarchical vision transformer model in this work. Patch embedding, Multi-Context Multi-Head Self-Attention (MHA), and Feed-Forward Network (FFN) with Layer Normalization are the three primary stages of the entire encoding process.

According to the patch embedding process, the input image is divided into $P \times P$ patches, and the obtained patches are then flattened and projected into an embedding space. This can be expressed as:

$$X_p = Reshape\left(ExtractPatches(I, P)\right) \in \mathbb{R}^{N\times D} \quad (17)$$

Where $N = \frac{H}{P} \times \frac{W}{P}$ represents the number of patches, D represents the embedding dimensions. The linear projection is applied to map these patches to a feature space where they can be represented as

$$Z_0 = W_p X_p + b_p, Z_0 \in \mathbb{R}^{N\times D} \quad (18)$$

Where W_p and b_p denote the learnable weights and bias parameters, respectively. Further, the obtained embeddings are processed through the Multi-Context Multi-Head Self-Attention (MHSA), which is used to capture the global dependencies across different regions of the image.

Unlike CNNs, which rely on local receptive fields, the proposed MHA allows the model to learn long-range dependencies and contextual relationships between different image regions. According to the process of MHSA, each token x_i is transformed into query (Q), key (K), and value (V) matrices, which are expressed as:

$$Q = W_Q X, K = W_k X, V = W_v X \quad (19)$$

Where $X \in \mathbb{R}^{N\times d'}$ represents the token sequence, $W_q, W_k, W_v \in \mathbb{R}^{d'\times d'}$. For these metrices, the self-attention can be computed as:

$$Attention(Q, K, V) = softmax\left(\frac{QL^T}{\sqrt{d'}}\right)V \quad (20)$$

In order to incorporate the multi-context awareness, the technique computes the attention at various hierarchical levels, which is expressed as:

$$H_l = \sum_{h=1}^{H} W_h . Attention(Q_h, k_h, V_h) \qquad (21)$$

Where H represents the number of attention heads and W_h represents the transformation weight for head h. Each level captures different granularities of attention (e.g., fine-grained local details vs. global context). Further, o structure the encoding, features are processed hierarchically, preserving both local and global information. This hierarchical processing of these attributes can be expressed as:

$$E_h = \sum_{l=1}^{L} W_l H_l \qquad (22)$$

Where L represents the levels, W_l It is used to dynamically adjust attention weights. A two-layer FFN is applied after attention:

$$FFN(X) = \sigma(XW_1 + b_1)W_2 + b_2 \qquad (23)$$

Further, Layer normalization and residual connections are used to ensure stable learning:

$$Z' = LayerNorm\left(Z_0 + Dropout\big(MHA(Z_0)\big)\right) \qquad (24)$$

$$Z_{final} = LayerNorm(Z' + Dropout(FFN(Z') \qquad (25)$$

This block improves feature representation and gradient flow, allowing the model to preserve contextual relationships effectively. The final encoder representation can be given as

$$E = LayerNorm(E_h) \qquad (26)$$

3.5. Decoder module

This section discusses the proposed decoder model, which is the Normalized Auto-Regressive Fine-Tune Transformer Decoder to process the encoded data. The proposed decoder module consists of several blocks, such as input token embedding, Positional Encoding, Auto-Regressive Attention, and a final prediction layer.

In order to represent each word in the target caption as a dense vector in a continuous embedding space. This allows the model to learn semantic relationships between words.

$$E_w = W_e Y + b_e \qquad (27)$$

Where Y represents the sequence of caption words, W_e and b_e These are the trainable parameters. According to this process, each word in the caption is first converted into an index corresponding to its position in the vocabulary. These indices are then mapped to vectors using the embedding matrix. W_e. The embedding space allows the model to capture semantic similarities between words. Moreover, the bias term ensures that each embedding can be shifted for better alignment in the vector space. In the next phase, the technique focuses on applying positional encoding to add information about word order, which Transformers do not inherently capture. Moreover, the positional encoding helps to retain the positional information.

This can be expressed as:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/D}}\right) \qquad (28)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/D}}\right) \qquad (29)$$

$$X_{input} = E_w + PE$$

This process ensures that the model distinguishes word positions, crucial for meaningful caption generation. Later, an autoregressive attention mechanism is applied to ensure that each predicted word depends only on past words, preventing information leakage. This attention mechanism can be expressed as:

$$MaskedAttention\ (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{D_k}} + M\right)V \ (30)$$

In this stage, the attention model computes attention scores using dot products between queries and keys, and $\frac{1}{\sqrt{D_k}}$ represents the scaling factor, which is used to stabilize the gradient factor. Later, this module uses Layer Normalization and Residual Connection to stabilize training and improve gradient flow by normalizing the input across the feature dimensions. The normalization layer is presented as

$$LayerNorm(x) = \frac{x-\mu}{\sigma}\gamma + \beta \qquad (31)$$

Where x is the input to the normalization layer, μ and σ are the standard deviation of the input, γ and β are the trainable and shifting parameters. The output of the attention layer is added back to the input (residual connection) to preserve information. Later, Layer Normalization is applied to stabilize and accelerate training convergence by ensuring that the input to each layer has a mean of zero and a variance of one.

3.5. Final Prediction Layer

Finally, a prediction layer is incorporated, which predicts the following word in the sequence. Mainly, it generates the probability distribution over the vocabulary for the next word in the sequence.

It is expressed in Equation 32 as follows:

$$P\big(y_t \big| y < t, z_{final}\big) = softmax\ (W_0 X_t + b_0)\text{words} \qquad (32)$$

Where y_t represents the predicated word at a given time stamp t, $y < t$ represents all previous, Z_{final} encoded image feature, w_0 and b_0 trainable parameters of the output layer and X_t represents the input to the final prediction layer.

3.6. Nucleus Sampling and Reranking Models

Several methods have worked on predicting the image caption, but the traditional methods have not considered the diversity, which affects the accuracy. Moreover, the existing methods do focus on reranking of the produced captions; therefore, these methods suffer from repetitive captions. Therefore, in this approach, generating diverse and relevant captions is crucial for enhancing user experience and model performance. Two advanced techniques, RoBERTa Reranking and Nucleus Sampling, are employed to achieve this. RoBERTa Reranking improves the relevance and fluency of generated captions by scoring and sorting them. Similarly, the Nucleus Sampling enhances diversity by selectively sampling from the most probable words, avoiding repetitive or generic outputs.

3.6.1. Nucleus Sampling

Nucleus Sampling selects words from the smallest possible set whose cumulative probability exceeds a threshold p. This approach maintains diversity while ensuring relevance. According to this process, for a given set of logits, the probabilities are calculated using the softmax function:

$$P(w_i) = \frac{e^{z_i}}{\sum_j e^{z_j}} \tag{33}$$

Later, sort the probabilities in descending order and compute the cumulative probability as follows $P(w_{i1}) \geq P(w_{i2}) \geq \cdots \geq P(w_{iN})$ and $C_k = \sum_{j=1}^{k} P\left(w_{ij}\right)$, respectively. Based on this, define the nucleus set. V_p as the smallest set where the cumulative probability is higher than the threshold:

$$V_p = \{w_{i1}, w_{i2}, \dots, w_{ik}\}\ such\ that\ C_k \geq p$$

Further, use this nucleus set and randomly sample the next word from this nucleus set as follows.

$$w_{t+1} \sim P\left(w \middle| V_p\right) \tag{34}$$

3.6.2. RoBERTa Reranking

The generated captions are then fed to a pre-trained RoBERTa model. This model computes the scores of each caption based on contextual relevance and fluency. The obtained captions are ranked according to these scores, and the highest-ranked captions are selected as the final output. According to this process, the first phase is to perform Tokenization and Encoding on the generated caption as:

$$X_i = RoBERTa\_Tokenizer(C_i) \tag{35}$$

The encoded sequence is then fed to the RoBERTa model to obtain the logits as follows:

$$L_i = RoBERTa(X_i) \tag{36}$$

Further, the relevance score is computed using softmax of logits, and relevance and irrelevance are estimated as a binary class to obtain the final class probability as

$$S_i = Softmax(L_i) \tag{37}$$

$$R_i = S_i[1] \tag{38}$$

Based on these relevance scores, the captions are ranked to obtain the reranked output as follows:

$$Rank(C) = argsort(R) \tag{39}$$

The proposed architecture overcomes the limitations identified, its performance is proven and confirmed to work correctly within the frames of systematic experimentation. Thus, the following section introduces the experimental design, data, and performance measures applied to evaluate the framework performance of the proposed framework.

The second important step after standardizing and preprocessing the input data is removing meaningful visual representations. Therefore, the following subsection explains the process of extracting features using the InceptionV3 architecture. In spite of the fact that feature extraction provides salient visual patterns, they need to be converted into semantically rich representations to be effectively captioned. This conversion is attained by proposing the encoder-decoder architecture as highlighted in the subsequent subsection.

Although the encoder concentrates on the modeling of hierarchical and global visual dependencies, the caption generation relies on sequential decoding, which is effective. Thus, the following sub-section discusses the normalized auto-regressive Transformer decoder that was applied to produce coherent and context-sensitive captions. The model should also predict the best linguistic output at every time step after the contextual representations have been decoded. The rest of the prediction layer, which produces word-level probability distributions, is described in the following sub-section. Even though probabilistic prediction can be used to generate captions, maximum-likelihood decoding may be used to generate captions, but can usually lead to repetitive captions. To overcome this shortcoming, the following subsection presents nucleus sampling and reranking based on RoBERTa to improve diversity and semantic relevance. Having the architectural elements and decoding tactics detailed to the end, the effectiveness of the proposed framework will need to be empirically tested. In the following section, the results of experimental findings and a comparative analysis of experimental results with those of the existing image captioning methods are provided.

4. Results and Discussion

This section presents the outcome of the proposed approach and compares the performance with the existing image captioning methods. The first subsection presents the details of the dataset used in this work, the following subsection presents the experimental setup for this work, later performance measurement parameters are discussed, and finally, experimental Analysis is presented.

4.1. Dataset Details and Experimental Setup

The given strategy is justified by publicly available Flickr8k [23] and Flickr 30k dataset [24]. It was simulated on a system with an NVIDIA RTX 2030 graphics card with 6 GB VRAM, 16GB RAM, an Intel Core i5 processor, and a Windows platform. This hardware setup offered enough computational resources to be able to train and fine-tune Deep Learning Models, and used the acceleration ability of the GPU to cut back on computation time and effectively process extensive batch training. The software environment contained Python as the programming language, with the leading deep learning frameworks being TensorFlow 2.15.0 and PyTorch. Subject to this, the Normalized Auto-Regressive Fine-Tune Transformer and Nucleus Sampling are implemented with the help of TensorFlow, and the reranking and fine-tuning of the RoBERTa model with the help of PyTorch. Pre-trained models such as RoBERTa and BERT were seamlessly integrated with the Hugging Face Transformers library and were able to handle tokenization and embedding, and can be executed with ease.

The simulation used InceptionV3 to extract features, which overcame spatial hierarchies of visual patterns that were at high levels of the input images, which were resized to (256, 256, 3). Such features were then decoded with sequential word tokens with the Normalized Auto-Regressive Fine-Tune Transformer, taking advantage of positional encoding and auto-regressive attention to preserve the word order and contextual relevance: creativity and relevance. The Nucleus Sampling was also applied to increase the diversity and quality of generated captions by sampling the most probable mass (p = 0.9). Reranking of captions generated was then done with the help of RoBERTa to make them semantically consistent and pick the most contextually suitable sequences.

The training was performed with AdamW optimizer, a learning rate of 2×10 5 and epsilon 10 -8, and Linear Scheduler with Warmup to stabilize the training process by varying learning rates. The data were pairs of images and captions, which were preprocessed with the help of BERT and RoBERTa tokenization and trained during 20 epochs with a batch size of 8. The simulation architecture allowed effective testing with the state-of-the-art models and generated captions of images of high quality with increased diversity and contextual accuracy. Combining the strong ability to extract features of InceptionV3, the sequential prediction model of the Transformer model, and the reranking enhancement feature of RoBERTa, this method proved to be effective in solving the problems of image captioning, generating contextually and semantically relevant captions. A full transparent description of the datasets, training procedure, as well as the hyperparameter settings used in the research, is provided in the following coherent configuration Table 2.

Table 2. Comprehensive dataset, training, and hyperparameter configuration

Category	Parameter	Specification
Dataset Details	Dataset Name	Flickr8k, Flickr30k
	Number of Images	8,000 (Flickr8k), 31,783 (Flickr30k)
	Captions per Image	5
	Total Captions	40,000 (Flickr8k), 158,915 (Flickr30k)
	Data Split	Train: 75%, Validation: 12.5%, Test: 12.5%
	Image Resolution	256 × 256 × 3
Image Preprocessing	Resizing	Uniform resizing to fixed input dimensions
	Normalization	Channel-wise mean and standard deviation normalization.
	Augmentation	Random cropping, horizontal flipping, rotation, and color jittering
Text Preprocessing	Tokenization	Word and subword tokenization
	Stopword Removal	Removal of non-informative tokens
	Lemmatization	Conversion to base word forms
	Vectorization	Pre-trained word embeddings
Feature Extraction	Visual Encoder	InceptionV3 (pre-trained)
	Feature Type	Global average pooled visual features.
Model Architecture	Encoder	Hierarchical Vision Transformer
	Decoder	Normalized Auto-Regressive Fine-Tuned Transformer
	Attention Mechanism	Multi-Context Multi-Head Self-Attention
Decoding Strategy	Decoding Method	Auto-regressive decoding

	Nucleus Sampling (p)	0.9
	Maximum Caption Length	20 tokens
Reranking Module	Language Model	RoBERTa (pre-trained)
	Selection Criterion	Highest semantic relevance score
Training Configuration	Optimizer	AdamW
	Learning Rate	1×10^{-4}
	Epsilon	1×10^{-8}
	Learning Rate Scheduler	Linear Scheduler with Warmup
	Batch Size	8
	Epochs	20
	Loss Function	Cross-Entropy Loss
	Dropout Rate	0.1
Implementation Setup	Frameworks	TensorFlow 2.15.0, PyTorch
	Supporting Libraries	Hugging Face Transformers
	Hardware	NVIDIA RTX 2030 GPU (6 GB VRAM), 16 GB RAM
	Operating System	Windows

The values of the architectural choices, the preprocessing steps, and the values of optimization are selected successfully, which ensures the stability of training, the reproducibility, and the high-performance evaluation. These experimental design choices together would make the plausibility and the technical soundness of the structure of the suggested image captioning.

4.2. Performance Measurement Parameters

Four various metrics are used in the performance of the proposed model, and they are METEOR (Metric for Evaluation of Translation with Explicit Ordering), ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence), BLEU Score (Bilingual Evaluation Understudy), and CIDEr Score (Consensus-based Image Description Evaluation).

CIDEr: CIDEr is used to evaluate an agreement between a generated caption and various reference captions, and it is perfect to evaluate picture captioning. To avoid such a tendency of over-punishment of innovative descriptions, it emphasizes n-gram matching with consideration of the phrase rarity. To reduce the effects of the popular words, it employs a Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme to determine the similarity of n-grams between the generated caption and the reference captions.

$$CIDEr = \frac{1}{N}\sum_{n=1}^{N} g_n . r_n \tag{40}$$

In which N is the size of the n-grams used, g n is the TF-IDF vector of n-grams in the generated caption, and r n is the TF-IDF vector of n-grams in the reference captions. BLEU is a precision-based measure that is used to examine the count of n-grams generated by captions that exist in the reference captions, and short captions are punished to prevent brevity bias. It compares the accuracy of n-grams (usually 1-4) in the obtained text to reference texts and even uses a brevity penalty to discourage extremely short sentences.

$$BLEU = BP.\exp(\textstyle\sum_{n=1}^{N} w_n \log p_n) \tag{41}$$

Where $BP = \begin{cases} 1, if\ c > r \\ \exp\left(1 - \frac{r}{c}\right), if\ c \le r \end{cases}$ c = length of the generated caption, r = length of the reference caption closest in length, w_n = weight for n-grams, p_n = precision for n-grams.

ROUGE-L: ROUGE-L emphasizes fluency and sentence structure preservation while concentrating on memory by comparing the generated and reference captions' Longest Common Subsequence (LCS). Taking into account the word order, it calculates the Longest Common Subsequence (LCS) to quantify sentence-level fluency.

$$ROUGE - L = \frac{(1+\beta^2).P.R}{R+\beta^2.P} \tag{42}$$

Where P and R are the precision and recall, and β is the balancing factor for P and R.

METEOR: It evaluates text similarity by aligning words using stemming, synonyms, and paraphrase matching, rewarding recall and penalizing word order differences. It focuses on recall, precision, and synonym matching, making it robust for evaluating semantic similarity.

$$METEOR = F_{mean}.(1 - Penalty) \tag{43}$$

4.3. Comparative Analysis

The performance of the proposed approach is compared with the state-of-the-art deep learning-based methods of image caption generation. Tables 3 and 4 demonstrate the obtained performance for the Flickr8k and 30k image datasets. The suggested model for image captioning on the Flickr8k dataset significantly improved performance, as shown by the comparison study in Table 1. BLEU scores (B1, B2, B3, B4), ROUGE L, METEOR, and CIDEr are among the evaluation metrics shown in the table. The suggested model achieves high BLEU scores across all n-grams, as shown by B1: 0.985, B2: 0.991, B3: 0.995, and B4: 0.991, indicating better precision at capturing unigrams to four-grams in comparison to previous models. Once the datasets and experimental setup have been defined, the required evaluation metrics will be needed to estimate caption quality objectively. The second sub-section presents the quantitative measures of measuring linguistic accuracy and semantic alignment. Although evaluation metrics offer numerical performance indicators, they can be better explained by comparing them with the existing methods. In this connection, the subsection below gives a comparative analysis in detail on benchmark datasets.

Table 3. Comparative analysis for the flickr8k dataset

Reference	B1	B2	B3	B4	ROUGE L	METEOR	CIDEr
[25]	0.579	0.383	0.245	0.160	NA	NA	NA
[26]	0.690	0.471	0.324	0.219	0.502	0.203	0.507
[27]	0.601	0.414	0.274	0.181	0.433	0.183	0.452
[28]	0.634	0.400	0.287	0.151	NA	NA	NA
[29]	0.589	0.335	0.263	0.148	NA	NA	NA
[30]	0.603	0.360	0.220	0.122	NA	NA	NA
[31]	0.674	NA	NA	0.243	0.448	0.215	0.636
[32] Ensemble	0.728	0.495	0.323	0.208	0.432	0.235	0.604
Proposed Model	0.985	0.991	0.995	0.991	0.99	0.99	0.989

Likewise, the ROUGE L (0.99) and METEOR (0.99) reported by the suggested method show that the generated captions are more semantically aligned with the reference captions. This emphasizes improved sentence fluency, phrase structure, and word choice. Furthermore, the model's remarkable alignment with human judgment is demonstrated by its CIDEr score of 0.989, which highlights how well it generates descriptive and contextually accurate captions. Likewise, we expanded on this comparison study and contrasted the results of the suggested model for the Flickr30k dataset. The quantitative Analysis proves that the proposed framework performs very well according to the traditional image captioning indicators. However, to see the actual effects and durability of the framework in practice, a more comprehensive analysis is required. The fact that metric values are reported does not go further to describe how or why the model is performing better. Thus, other dimensions of Analysis are added to give a holistic analysis of the offered architecture in terms of the contribution of the components, caption diversity, and qualitative behavior in actual scenarios.

Table 4. Comparative analysis for flickr30k dataset

Reference	B1	B2	B3	B4	ROUGE L	METEOR	CIDEr
[25]	0.573	0.369	0.240	0.157	NA	NA	NA
[33]	0.647	0.460	0.324	0.230	0.189	NA	NA
[34]	0.649	0.462	0.324	0.224	0.451	0.194	0.472
[35]	0.677	0.494	0.354	0.251	0.204	NA	0.531
[36]	0.666	0.484	0.346	0.247	0.467	0.202	0.524
[26]	0.689	0.468	0.319	0.220	0.487	0.191	0.428
[37]	0.695	0.463	0.341	0.232	0.451	0.302	0.486
[38]	0.647	0.456	0.320	0.224	0.449	0.197	0.467
[31]	0.671	NA	NA	0.233	0.443	0.204	0.645
[39]	0.694	0.498	0.355	0.254	0.538	0.251	0.469
Ensemble [32]	0.798	0.561	0.387	0.269	0.443	0.213	0.565
Proposed model	0.985	0.992	0.986	0.985	0.991	0.9850	0.991

In order to assess the importance of every architectural part, a component-wise ablation is performed. This evaluation evaluates in a methodical manner the effect of the Vision Transformer encoder, nucleus sampling approach, and reranking module based on RoBERTa by successively mobilizing each component. The findings, presented in Table 3, show that the Vision Transformer is much more effective at representing the context, with long-range spatial dependencies. In contrast, nucleus sampling is much less effective at representing linguistic diversity because it relies too heavily on high-probability tokens. Quality of the captions is further improved with the incorporation into the RoBERTa reranking framework that puts emphasis on semantically sound and context-related sentences. The complementary character of these components and the overall improvement is not facilitated by one of the modules, which is confirmed by the progressive performance improvement that is observed in all evaluation metrics.

The Proposed Model outperforms all other approaches across all evaluation metrics, achieving the highest BLEU scores (B1-B4), ROUGE L, METEOR, and CIDEr. Compared to the existing ensemble [26] approach, the Proposed Model demonstrates significant improvements: BLEU-1 improves from 0.798 to 0.985, indicating better unigram precision. BLEU-4 increases from 0.269 to 0.985, reflecting superior fluency and contextual accuracy. CIDEr rises from 0.565 to 0.991, showing enhanced semantic alignment with human-generated captions. Further, the qualitative Analysis is also presented to demonstrate the output of the proposed model without reranking and with a reranking mechanism. The following figure depicts the outcome of these two models.

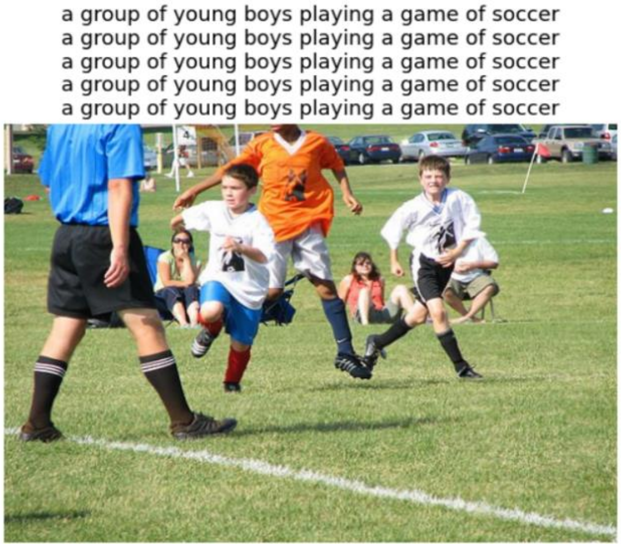
a group of young boys playing a game of soccer
a group of young boys playing a game of soccer
a group of young boys playing a game of soccer
a group of young boys playing a game of soccer
a group of young boys playing a game of soccer

Fig. 3 Repetitive caption generation for a soccer scene

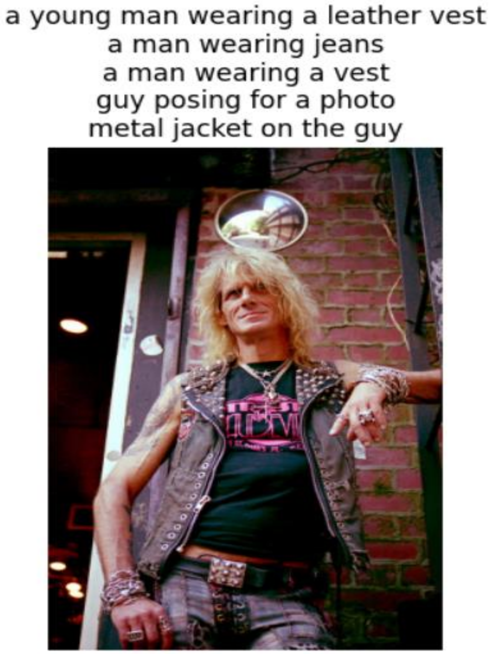

Fig. 4 Ambiguous caption generation for a portrait image

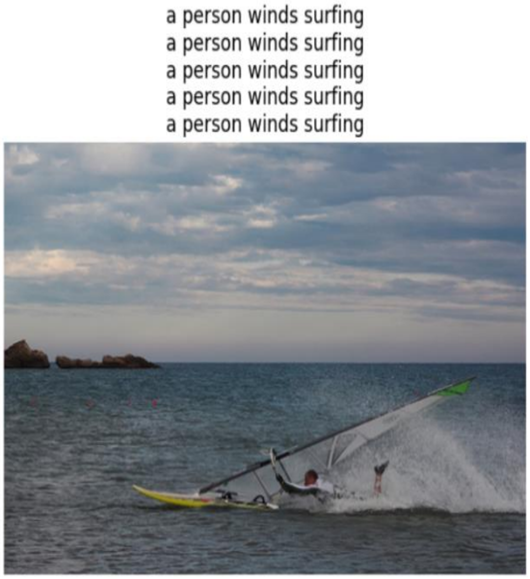

Fig. 5 Repetitive captioning in action-based water sports scene

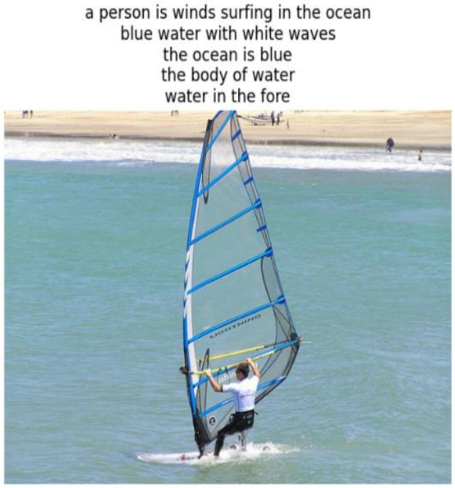
a person is winds surfing in the ocean
blue water with white waves
the ocean is blue
the body of water
water in the fore

Fig. 6 Context-aware captioning for ocean windsurfing scene

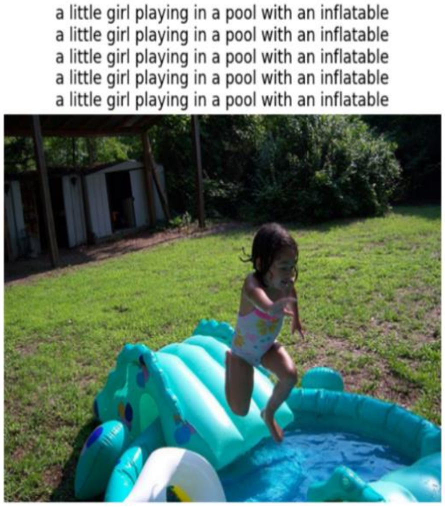
a little girl playing in a pool with an inflatable
a little girl playing in a pool with an inflatable
a little girl playing in a pool with an inflatable
a little girl playing in a pool with an inflatable
a little girl playing in a pool with an inflatable

Fig. 7 Redundant caption outputs for a child play activity

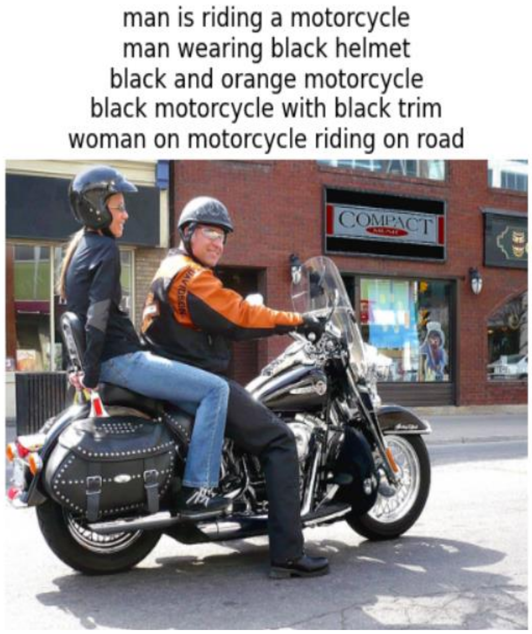
man is riding a motorcycle
man wearing black helmet
black and orange motorcycle
black motorcycle with black trim
woman on motorcycle riding on road
COMPACT

Fig. 8 Mixed-semantic caption generation for a motorcycle riding scene

Besides accuracy-based measures, the caption diversity and fluency are also evaluated in order to determine the expressive power of the proposed approach. Conventional methods of decoding are usually used to produce repetitive or too generic captions that restrict their application in real life. The proposed framework, by introducing nucleus sampling, has a better ratio of unique captions and longer and descriptive sentence structures. Repeatedly, in comparison to traditional transformer-based models, as shown in Table Y, the repetition rate is significantly lower, and semantic relevance is preserved. This trade-off between diversity and accuracy is especially relevant to those applications that rely on descriptive richness, like accessibility systems and content retrieval, where the user experience directly depends on the descriptive richness.

The qualitative Analysis of the results is also provided, complementing the numerical results. The captions of the samples that have been developed by the reranking mechanism and the ones that have not are compared to show the effects of the semantic refinement. The qualitative results show that the captions selected according to the RoBERTa-generated reranking module are better grammatically constructed, better connected to the action of the object, and more consistent with visual content. These remarks validate quantitative advantages and images that the planned model yields labels that are statistically better, besides being more perceptually instructive and human-like. Lastly, the gains are compared with the state-of-the-art practices to bring out the practical importance of the gains reported. Table Z shows the improvements in the main metrics (percentages) and shows an overall regular improvement in the performance in benchmark datasets. As opposed to the previous works that only address absolute scores, the current Analysis will rely on interpretability by evaluating the extent of the improvement that the suggested approach can bring. All these protracted analyses reinforce the empirical base of the research and establish that the suggested framework presents a significant contribution to the image captioning research instead of a slight increase in performance. Whereas comparative outputs confirm that the proposed framework is more effective, there is more to be done to perceive why performance improvements are attained. Thus, a more extended discussion is introduced below, with the emphasis put on component contributions, diversity, and qualitative behavior. The amount of the quantitative and qualitative analyses proves the effectiveness and strength of the offered image captioning framework. On the basis of these findings, the final part of the work is a summary of significant contributions and a prospective direction of future research.

5. Conclusion

This paper describes a novel structure of image captioning that successfully overcomes the drawbacks of traditional encoder-decoder structures, such as poor contextual knowledge, redundancy, and insufficient language variety. The suggested model maps intricate spatial relationships, creates consistent and contextually correct captions by combining a Vision Transformer encoder and a Normalized Auto-Regressive Fine-Tune Transformer decoder. Although the reranking based on RoBERTa ensures language fluency and applicability to context, Nucleus Sampling increases the variety. The proposed algorithm achieves higher scores on such significant evaluation criteria as CIDEr, BLEU, ROUGE-L, and METEOR, which suggests significant improvements in the quality of captions and the adequacy of the context. The results have created a precedent of image captioning and affirmed the effectiveness of our method in the creation of natural and comprehensive captions. Besides increasing the bar, this finding opens up more chances of future research, including the exploration of cross-modal interactions and advancement in interpretability. Besides furthering vision-language comprehension, the proposed architecture holds a lot of viable potential for application in assistive technologies, multimedia retrieval, and human-computer interaction.

Ethical Statement

No human subjects were used in this study; no human subjects are involved, and no personally identifiable human data were used. All experiments were carried out based on publicly available benchmark data, which contains non-sensitive visual objects and related annotations only used to conduct research. Since no human data were gathered, manipulated, or discussed, there was no need for institutional review board or ethics committee approval. The research is conducted in accordance with the general research ethics and data use.

References

[1] L. Ashok Kumar, and D. Karthika Renuka, *Deep Learning Approach for Natural Language Processing, Speech, and Computer Vision, Techniques and Use Cases*, 1st ed., CRC Press, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] V. Ajantha Devi, and Mohd Naved, *Dive in Deep Learning: Computer Vision, Natural Language Processing, and Signal Processing*, *Machine Learning in Signal Processing*, 1st ed., Chapman and Hall/CRC, 2021. [Google Scholar] [Publisher Link]

[3] Mohammad Mustafa Taye, "Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions," *Computers*, vol. 12, no. 5, pp. 1-26, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Shervin Minaee et al., "Image Segmentation using Deep Learning: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3523-3542, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Reza Azad et al., "Medical Image Segmentation Review: The Success of U-Net," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10076-10095, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[6] Yun Yang et al., "Two-stage Selective Ensemble of CNN via Deep Tree Training for Medical Image Classification," *IEEE Transactions on Cybernetics*, vol. 52, no. 9, pp. 9194-9207, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[7] Ahmed A. Elngar et al., "Image Classification based on CNN: A Survey," *Journal of Cybersecurity and Information Management*, vol. 6, no. 1, pp. 18-50, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[8] Akhilesh Kumar Sharma et al., "An Efficient Approach of Product Recommendation System using NLP Technique," *Materials Today: Proceedings*, vol. 80, pp. 3730-3743, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Suraj Bodapati et al., *Comparison and Analysis of RNN-LSTMs and CNNs for Social Reviews Classification*, Advances in Applications of Data-Driven Computing, pp. 49-59, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] Shashank Mohan Jain, *Introduction to Transformers for NLP*, With the Hugging Face Library and Models to Solve Problems, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[11] Chia Xin Liang et al., "A Comprehensive Survey and Guide to Multimodal Large Language Models in Vision-Language Tasks," *arXiv preprint*, pp. 1-115, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[12] Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar, "Deep Learning Approaches on Image Captioning: A Review," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1-39, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] Ramesh Sneka Nandhini, and Ramanathan Lakshmanan, "QCNN_BaOpt: Multi-Dimensional Data-based Traffic-Volume Prediction in Cyber-Physical Systems," *Sensors*, vol. 23, no. 3, pp. 1-16, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Bowen Xin et al., "A Comprehensive Survey on Deep-Learning-based Visual Captioning," *Multimedia Systems*, vol. 29, no. 6, pp. 3781-3804, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Qazi Anwar, and Ch.V.S. Satyamurty, "An Analysis on Recent Approaches for Image Captioning," *CVR Journal of Science and Technology*, vol. 26, no. 1, pp. 87-92, 2024. [Google Scholar] [Publisher Link]

[16] Jafar A. Alzubi et al., "Deep Image Captioning using An Ensemble of CNN and LSTM based Deep Neural Networks," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 4, pp. 5761-5769, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Samar Elbedwehy et al., "Efficient Image Captioning based on Vision Transformer Models," *Computers, Materials & Continua*, vol. 73, no. 1, pp. 1483-1500, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Akash Verma et al., "Automatic Image Caption Generation using Deep Learning," *Multimedia Tools and Applications*, vol. 83, no. 2, pp. 5309-5325, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Ahatesham Bhuiyan et al., "Enhancing Image Caption Generation through Context-Aware Attention Mechanism," *Heliyon*, vol. 10, no. 17, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Santosh Kumar Mishra et al., "Efficient Channel Attention based Encoder-Decoder Approach for Image Captioning in Hindi," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 3, pp. 1-17, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[21] Shourya Tyagi et al., "Novel Advance Image Caption Generation Utilizing Vision Transformer and Generative Adversarial Networks," *Computers*, vol. 13, no. 12, pp. 1-23, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[22] Huimin Lu et al., "Chinese Image Captioning via Fuzzy Attention-based DenseNet-BiLSTM," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 1S, pp. 1-18, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[23] Flickr 8k Dataset, Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/adityajn105/flickr8k

[24] Hsankesara, Flickr Image Dataset, Kaggle, 2018. [Online]. Available: https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset

[25] Andrej Karpathy, and Li Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3128-3137, 2015. [Google Scholar] [Publisher Link]

[26] Teng Jiang, Zehan Zhang, and Yupu Yang, "Modeling Coverage with Semantic Embedding for Image Caption Generation," *The Visual Computer*, vol. 35, no. 11, pp. 1655-1665, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[27] Amish Patel, and Aravind Varier, "Hyperparameter Analysis for Image Captioning," *arXiv preprint*, pp. 1-10, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[28] Harshitha Katpally, and Ajay Bansal, "Ensemble Learning on Deep Neural Networks for Image Caption Generation," *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, San Diego, CA, USA, pp. 61-68, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[29] J. Bineeshia, "Image Caption Generation using CNN-LSTM based Approach," *ICCAP 2021: Proceedings of the First International Conference on Combinatorial and Optimization*, Chennai, India, 2021. [Google Scholar]

[30] Zagon Bussabong et al., "Enhancing Image Caption Performance with Improved Visual Attention Mechanism," *ICIC Express Letters Part B: Applications*, vol. 16, no. 1, pp. 73-82, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[31] Yiwei Ma et al., "Towards Local Visual Modeling for Image Captioning," *Pattern Recognition*, vol. 138, pp. 1-32, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[32] Israa Al Badarneh, Bassam H. Hammo, and Omar Al-Kadi, "An Ensemble Model with Attention-based Mechanism for Image Captioning," *Computers and Electrical Engineering*, vol. 123, pp. 1-35, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[33] Quanzeng You et al., "Image Captioning with Semantic Attention," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 4651-4659, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[34] Kun Fu et al., "Aligning Where to See and What to Tell: Image Captioning with Region-based Attention and Scene-Specific Contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2321-2334, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[35] Jiasen Lu et al., "Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 3242-3250, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[36] Chen He, and Haifeng Hu, "Image Captioning with Text-based Visual Attention," *Neural Processing Letters*, vol. 49, no. 1, pp. 177-185, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[37] Tiago do Carmo Nogueira et al., "Reference-based Model using Multimodal Gated Recurrent Units for Image Captioning," *Multimedia Tools and Applications*, vol. 79, no. 41-42, pp. 30615-30635, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[38] Marimuthu Kalimuthu et al., "Fusion Models for Improved Image Captioning," *Pattern Recognition, ICPR International Workshops and Challenges*, pp. 381-395, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[39] Amr Abdussalam et al., "Numcap: A Number-Controlled Multi-Caption Image Captioning Network," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 4, pp. 1-24, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[40] Jianlin Su et al., "Roformer: Enhanced Transformer with Rotary Position Embedding," *Neurocomputing*, vol. 568, pp. 1-14, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[41] Shaowen Wang, Linxi Yu, and Jian Li, "LoRA-GA: Low-Rank Adaptation with Gradient Approximation," *arXiv preprint*, pp. 1-19, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[42] Yue Yang et al., "Remote Sensing Image Change Captioning using Multi-Attentive Network with Diffusion Model," *Remote Sensing*, vol. 16, no. 21, pp. 2024. [CrossRef] [Google Scholar] [Publisher Link]

[43] Yongshuo Zhu et al., "Semantic-CC: Boosting Remote Sensing Image Change Captioning via Foundational Knowledge and Semantic Guidance," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1-16, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[44] Shuai Bai et al., "Qwen2.5-vl Technical Report," *arXiv preprint*, pp. 1-23, 2025. [CrossRef] [Publisher Link]

[45] Yuduo Wang, Weikang Yu, and Pedram Ghamisi, "Change Captioning in Remote Sensing: Evolution to SAT-Cap-A Single-Stage Transformer Approach," *arXiv preprint*, 1-18, 2025. [CrossRef] [Google Scholar] [Publisher Link]