Attention-Enhanced Efficient-Net for Feature Extraction in Transformer-Based Image-to-Text Generation

1st Anjali Sharma
FET, Gurukul Kangri Deemed to be University

2nd Dr. Mayank Aggarwal
Department of Computer Science and Engineering

***Abstract*—Recent years have seen tremendous progress in computer vision, especially in areas such as image classification and object identification. Nowadays, image captioning is one of the recent and growing research problems. There is an ongoing need for systems that are more precise and efficient despite the existence of extant solutions. The primary goal of this work is to develop an encoder-decoder architecture that incorporates three distinct attention mechanisms for utilization in an automated image captioning system. This work utilizes the COCO-2017 dataset, images and reference captions. EfficientNetB0, enhanced with Global attention, is employed to resize and analyze images in order to extract features. The dataset is bifurcated into two parts: one with 20,000 images for testing purposes and the other with 90,000 images for training. This work provides a model for generating image description that combines features extracted using EfficientB0 with text creation utilizing a transformer based encoder-decoder with Multihead Attention and Token level Adaptive Attention. The proposed model is assessed using the BLEU, ROUGE, and CIDEr metrics, resulting in a high-performance score of 0.8326, ROUGE-1 of 0.9422, ROUGE-2 of 0.9003, and CIDEr of 0.8563. The potential of attention-enhanced Transformer-based algorithms to generate correct and coherent image captions is demonstrated in this work, with an aggregate test accuracy of 79.87 percent. The results demonstrate that the model effectively caught key visual aspects since the reference captions and the generated captions showed a high level of agreement.**

***Index Terms*—Automated Image Caption Generation System, Global Attention, EfficientNetB0, Token Level Adaptive Attention**

I. Introduction

Computer Vision and image processing have advanced a lot in the last couple of years, especially with regard to object recognition and image classification [1] [2]. The benefits of automatically producing natural images and full descriptions are more significant in the following areas: text-based image retrieval, data access for blind users, healthcare image title descriptions, and news image captions [3]. There are important research implications for both theory and practice in this image captioning application. Since AI technology has advanced, image captioning has become a challenging but practical endeavor [4].

The goal of automatic picture captioning, a difficult computer vision problem, is to provide rich material and descriptions that are comprehensible to humans for supplied images [5]. The proliferation of digital images has forced us to cope with a wide variety of online image resources, such as news stories, ads, blogs, and the like [6] [7]. Most photographs don't have a description, which makes it hard for users to understand them, and even when a description is included, it requires a lot of work to manually confirm that it matches the image. Therefore, automatic picture captioning techniques are needed to characterize the content of photos due to the growing volume of images [8].

Although deep learning models have achieved impressive results, they often produce vague or overly generic captions. This limitation arises from encoding all visual information into a single vector, which can lead to inadequate representation of detailed image content [9], [10]. Many studies have used the Attention Mechanism (AM) in encoder-decoder architectures to address these issues; this mechanism uses an attention algorithm with target picture cues to give visual data more weight in the encoder design. Consequently, attention aids the technique in focusing on the crucial regions of the image. To accomplish this challenge, several strategies were put forth, including deep learning. [11], transformers have provided solutions for a variety of picture captioning issues. For example, the transformer learns long-range relations to attention to complete sequences, while recurrent networks focus on short-term context. Transformers aid in the isolation of crucial features by encoding the object area and then converting it to a vector representation, allowing for the simultaneous processing of sequences [12].

A. Novelty and Motivation

This paper proposes a new work in constructing image captioning using an Autoregressive transformer-based Encoder-Decoder model with Attention-enhanced EfficientNet-B0 for feature extraction by triple attention techniques, namely Global, multi-head and token level adaptive attention. The work aims to achieve a higher quality of captions than ex-

isting methods by optimizing the current encoder and decoder architecture. Furthermore, the work extends to exploit the generated high-quality captions to act as a prompt for the transformer-based text generation mechanism to produce high-quality story-like narration aligned with the given image. Conducted using a COCO-2017 dataset and assessing the model's performance against benchmark metrics, also joins the ongoing discussion around improving image captioning systems. This research advances image understanding systems to improve visual data perception and interaction, benefiting areas like accessibility, content delivery, and user experience. It also integrates image captioning with deep learning methods such as text generation and storytelling, reducing dependence on reference texts for more autonomous systems.

B. Aim and Contribution

The key contributions of this work are as follows:

- Efficient Feature Extraction with Global Attention
- Robust Preprocessing Pipeline
- Performance Optimization via Compound Scaling
- Advanced Decoder with Dual Attention Mechanisms
- Comprehensive Evaluation Metrics
- Interactive GUI with Caption and Story Generation

II. Literature Review

To gain a better understanding of the previous work and approaches that contribute to building image captioning systems, this section presents the literature review on automated Image captioning systems.

[13] have introduced an RNN method that uses LSTM to create image-based natural language. The dataset they use to train their machine comprises 8,000 photos with 37611 captions. Characteristic extraction from images is another usage of VVG16. When performance is finally assessed, the results indicate a 66% accuracy rate and BLEU-(1 to 4) scores of 0.40, 0.18, 0.11, and 0.03 respectively.

[14] suggest a DL model that creates captions and characterizes images using machine translation and computer vision. Visual objects and their relationships are correctly identified and labeled by the model. The creation and operation of neural networks are also examined in this paper. A BELU Score of 69.8 is attained by the suggested model.

[15] recently displayed an operation of 3 separate CNN models and highlighted the exceptional accuracy achieved by each: Xception, VGG-16, and ResNet50. The Flickr_8k dataset, which includes 8091 pictures, is utilized in their proposed project. This is then used to construct sentences. Comparing the BLEU scores—0.79 for Xception Model, 0.75 for VGG-16, and 0.84 for ResNet50—the three systems provide high-quality results and captions. The best network for feature extraction and categorization was found to be ResNet50, which achieved 84% accuracy in captions over 50 epochs. It also makes it easier to solve the vanishing gradient issue.

[16] An automatic description of an image is produced by applying deep learning and NLP through object detection and text generation. The architecture averaged a BLEU score of 51.77 and used a CNN encoder and an RNN decoder within a dense attention model.

[17] offer an encoder-decoder architecture that could result in grammatically sound image captions. The model uses LSTM for decoding and VGG16 Hybrid Places 1365 for encoding. All common measures, including BLEU, are used to evaluate the model. The suggested model achieved BLEU-1 to 3 scores of 0.603774, 0.388514, and 0.244706 on the Flickr 8k dataset respectively, according to experimental results. Comparing the proposed strategy to the advanced methods, a notable performance was obtained.

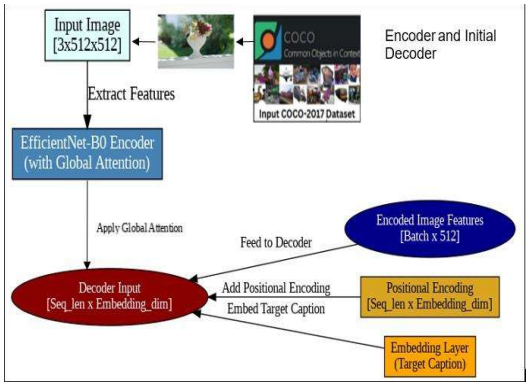
[18] work contributes by summarizing several methodologies, suggesting datasets to train and test picture captioning models, and implementing a CNN and LSTM based model. The LSTM model uses the extracted image features from the CNN model to produce natural language text. A BLEU-one-gram score of 0.755367 is obtained when the model is tested on Flickr 8k.

[19] comprehensively investigate DNNs-based image caption production. The model can use CNNs, RNNs, and sentence synthesis to take in images and produce English sentences that describe what's in them. Based on the Flickr 8k dataset, which contains more than 8,000 photos, these models were developed. Natural languages used by humans are typically brief and to the point when describing a scenario.

[20] The suggested generative model employs a deep convolutional neural network (VGG-19) to produce the most relevant feature vectors from the images. The research starts with training the model on popular datasets like FLICKR-8K, and then it utilizes the BLEU score—which can range from 0 to 100—to check if the model is accurate. Using the BLEU score, they evaluate their model with four others.

III. Methodology

The proposed captioning approach combines CNN and Transformer architectures, beginning with the COCO-2017 dataset containing images and corresponding descriptions. Preprocessing involves contraction mapping, removal of stopwords and punctuation, and the inclusion of START/END tokens. Images are resized to 512×512 pixels and converted to tensors. An EfficientNetB0 encoder, integrated with a Global Attention module, extracts high-level spatial features and emphasizes critical image regions, producing a global feature vector. The dataset is split into 90,000 training and 10,000 testing images for model evaluation. The hybrid model employs an encoder-decoder Transformer with multi-head and token-level adaptive attention mechanisms. Each decoder layer includes self-attention, cross-attention, and feed-forward sublayers with residual connections, normalization, and dropout for enhanced learning stability. This system, termed the Hybrid Model, effectively merges EfficientNetB0-based feature extraction with Transformer-based text generation. Performance is validated using BLEU, ROUGE, and CIDEr scores, with the added capability of story generation from the produced captions.



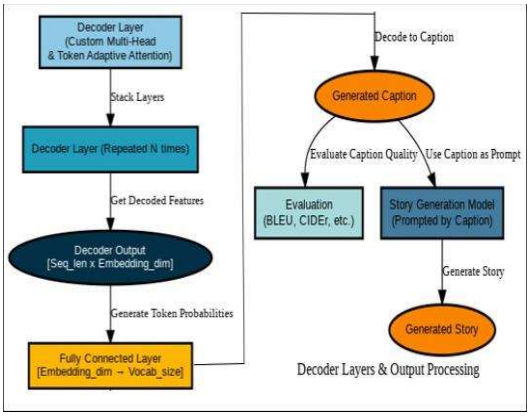


Fig. 1. Proposed System Flowchart for Automated Image Captioning System

Fig. 2. Images and Captions within the COCO-2017 Dataset

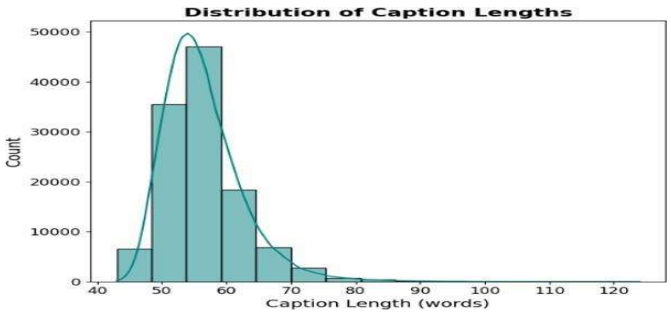


Fig. 3. Count Plot for Distribution of Caption Lengths in COCO-2017 Dataset

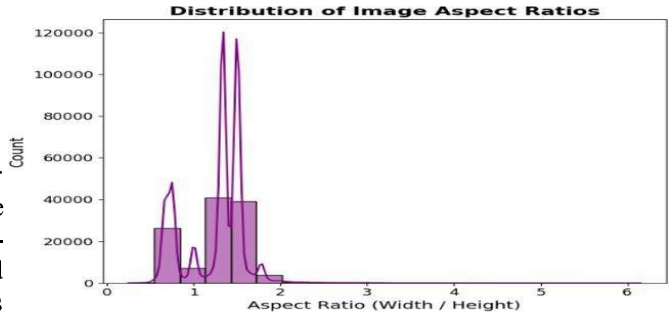


Fig. 4. Count Plot for Distribution of Aspect Ratio in COCO-2017 Dataset

A. Data preprocessing

Data preprocessing is critical for preparing raw data for analysis and modeling. In this work, several key steps were applied to ensure clean and effective inputs for image captioning. A contraction mapping dictionary was used to expand contracted words, improving readability. English stopwords were identified and removed, while extra spaces, punctuation, and inconsistent quotation marks were cleaned. Captions were converted to lowercase, tokenized, and framed with START and END tokens to define context. Images were resized to 512×512 pixels and converted to tensors to ensure compatibility with the model. Figure 2 displays the COCO-2017 dataset with the accompanying descriptions for each picture. Figure 3 displays the distribution of caption lengths in the dataset, shown as a count plot. Figure 4 shows the distribution of aspect ratios (width/height) in the dataset using a count plot.

B. Image Feature Extraction: EfficientNetB0 Model

For visual feature extraction, this work employs the **EfficientNetB0** model, a convolutional neural network (CNN) known for delivering high accuracy with low computational demand. The model architecture is composed of three major components: a *Conv stem*, a residual *body* consisting of MBConv blocks, and a *Conv head*. The MBConv blocks utilize *depthwise separable convolutions* to reduce computational complexity, along with *squeeze-and-excitation (SE)* layers that dynamically recalibrate channel-wise feature responses to enhance focus on informative regions [21]. A structural overview of EfficientNetB0 is presented in Figure 5.

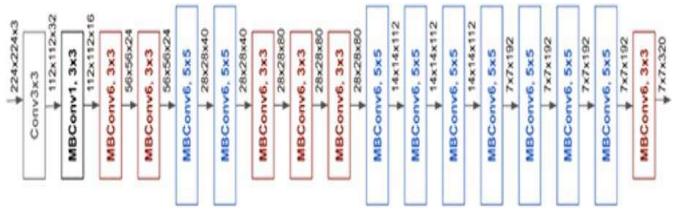


Fig. 5. Architecture of EfficientNet-B0 Model

The **STEM** module begins with a convolutional layer using

32 filters and a 3×3 kernel with a stride of 2. It is followed by batch normalization and ReLU6 activation, which together help in downsampling the input and extracting initial low-level visual patterns.

The **BODY** comprises stacked MBConv blocks, which separate spatial and channel convolutions to minimize parameter usage. Integrated squeeze-and-excitation blocks allow dynamic reweighting of feature channels, improving the model's representational efficiency.

The **HEAD** includes a global average pooling layer that compresses spatial dimensions into single values per channel. The original classification head (with Softmax activation) is replaced in this work; instead, the resulting feature vector serves as an image embedding input to the caption generation model.

One of the notable advancements in EfficientNetB0 is its **compound scaling** approach, which proportionally increases the network's depth, width, and input resolution in a balanced manner, guided by a unified scaling formula:

$$\text{Width} \times \text{Depth}^2 \times \text{Resolution}^2 \approx \text{Constant} \quad (1)$$

This balanced scaling improves performance without significantly increasing computational cost, providing a more efficient model compared to traditional CNN architectures.

To enhance the spatial awareness of the extracted features, a **Global Attention** mechanism is applied to the output of EfficientNetB0. The feature map x with dimensions $[B, C, H, W]$ is reshaped to $[B, H \times W, C]$, flattening the spatial dimensions. A learnable matrix W_{attn} generates attention scores for each spatial position, which are then normalized using a softmax function. These scores represent the importance of each spatial location and are used to reweight the feature map accordingly. The result is a globally weighted feature vector g, which highlights the most relevant regions of the image for downstream tasks such as caption generation.

Additionally, to complement the visual attention mechanisms, this work incorporates a **Token Level Adaptive Attention** mechanism during the decoding phase of caption generation. This module learns an adaptive weight matrix that dynamically assigns significance to each token in the input sequence. The resulting attention scores, normalized via softmax, form a probability distribution that emphasizes tokens of higher contextual relevance. This not only refines the importance of individual tokens but also enhances semantic coherence across the generated text. By applying token-level attention before multi-head attention, the model effectively blends fine-grained token focus with broader contextual relationships, leading to more fluent and accurate captions.

By integrating EfficientNetB0's compound scaling and architectural strengths with Global Attention for spatial focus and Token Level Adaptive Attention for linguistic refinement, the system achieves efficient and high-quality image feature extraction suitable for advanced image captioning applications.

C. Text Generation Using Transformer

The proposed image captioning system employs a hybrid Encoder-Decoder architecture that integrates EfficientNetB0 for visual feature extraction with a Transformer-based decoder for text synthesis. This design effectively captures both visual semantics and linguistic context, enabling coherent and accurate captioning.

EfficientNetB0, adapted from its original classification role, is modified to output a 512-dimensional embedding vector, encapsulating essential image features. This embedding serves as the input to the Transformer decoder.

The decoder utilizes multi-head attention to enable the model to attend to multiple aspects of both the image embedding and the generated caption sequence concurrently, enhancing contextual understanding during generation. Token-level adaptive attention further enhances this process by dynamically assigning weights to each token, emphasizing contextually important words and improving semantic coherence. This adaptive attention precedes multi-head attention, enabling more refined token relevance modeling.

Each decoder layer includes self-attention, cross-attention, and a feed-forward network, supported by residual connections, layer normalization, and dropout to ensure stability and generalization. Positional encoding is added to input tokens to preserve word order, which is essential for maintaining sentence structure in the generated captions.

The combined Encoder-Decoder model processes input images to generate captions sequentially, token by token. Key parameters include:

tgt_vocab_size: size of the vocabulary
d_model = 512: embedding dimensionality
num_heads = 8: number of attention heads
num_layers = 6: number of decoder layers
d_ff = 2048: feed-forward network size
max_seq_length: maximum length of output sequence
dropout = 0.1: regularization
num_epochs = 10: total training epochs

This architecture, evaluated using BLEU, ROUGE, and CIDEr metrics, demonstrates strong capability in generating fluent, semantically aligned captions. Its combination of EfficientNetB0 with advanced attention mechanisms offers a high-performing solution for deep learning-based image captioning.

D. Parameters for Training the Model

Table I illustrates the parameters used for training the model.

E. Model Evaluation

Model evaluation plays a pivotal role in validating machine learning systems. In this work, the proposed model's effectiveness was rigorously measured using a combination of evaluation metrics, including CIDEr, BLEU, ROUGE, and accuracy. This multi-metric approach ensures a thorough and dependable assessment of the model's performance across linguistic precision, semantic relevance, and classification reliability.

TABLE I
TRAINING PARAMETERS AND CONFIGURATIONS

Component	*Parameter*	*Value*
Loss Function	Criterion	Cross Entropy Loss
Optimizer	Optimizer Type	Adam
Optimizer	Learning Rate	0.0001
Optimizer	Betas	(0.9, 0.98)
Optimizer	Epsilon	1e-9
Scheduler	Type	Cosine Annealing
Scheduler	Iterations	10
Epochs	Training Duration	10
Hardware	Device Used	GPU
Monitoring	Metrics	Loss, Accuracy

- **BLEU:** A precision-oriented evaluation metric that quantifies the degree of overlap between n-grams in the generated and reference captions. BLEU-N (for N = 1 to 4) computes the geometric mean of n-gram precisions, where higher values indicate stronger alignment with human-authored descriptions.
- **ROUGE:** A recall-focused metric that assesses the match between candidate and reference sequences by examining unigram, bigram, and subsequence overlaps. ROUGE-1 captures individual word matches, ROUGE-2 evaluates word pairings, and ROUGE-L measures the longest common subsequence, considering word order and structure [22].
- **CIDEr:** Tailored for image captioning tasks, CIDEr evaluates semantic similarity by applying TF-IDF weighting to n-grams and computing the cosine similarity between resulted and given captions. It emphasizes informative content words, making it more sensitive to semantic richness than traditional n-gram methods [23].

IV. EXPERIMENTAL RESULTS ANALYSIS

The outcomes of a suggested DL model for automatic picture captioning are detailed in this section. The research has been conducted using Jupyter Notebook on a Windows 11 HP PC equipped with a 512 GB SSD, 16 GB RAM, and an AMD Radeon RX 6600 GPU.

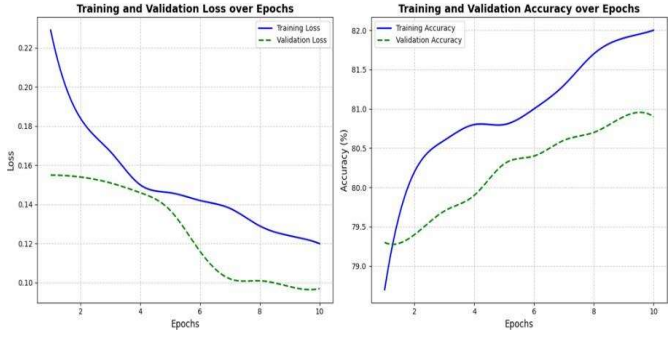


Fig. 6. Line Graph for Image Captioning Model Training/validation Accuracy and Loss

Figure 6 illustrates the progression of both training and validation metrics for the Hybrid model over a span of ten epochs. The left graph shows a constant decline in training loss from 0.22 to 0.1 and in validation loss from just under 0.16 to around 0.12, indicating effective learning. The right graph illustrates accuracy improvement, with training accuracy increasing from 79% to over 82%, and validation accuracy rising from 79% to approximately 80.9%. These trends reflect consistent model improvement. Table 2 provides the qualitative results supporting the model's effectiveness.

The average results for the three metrics BLEU, ROUGE, and CIDEr that are utilized to generate description for images are shown in Figure 7 and Table II. A horizontal bar graph is shown in the chart, and each measure has a score between 0 and 1. Notably high ROUGE-1 and ROUGE-2 scores—0.9422 and 0.9003, respectively indicate good recall and accuracy performance. An impressive CIDEr score of 0.8563 indicates that produced and reference captions are well aligned. BLEU scores show considerable n-gram overlaps in the produced captions; BLEU-1 is 0.8326, BLEU-2 is 0.6483, BLEU-3 is 0.5382, and BLEU-4 is 0.4919.

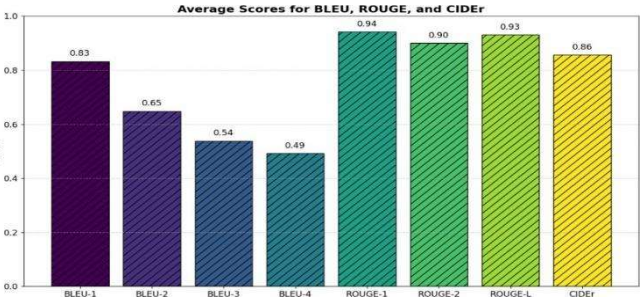


Fig. 7. Average Scores for BLEU, ROUGE and CIDEr Metrics in Image Caption Generation

TABLE II
AVERAGE SCORES OF PERFORMANCE METRICS FOR HYBRID MODEL

BLEU/CIDEr Scores				
B-1	B-2	B-3	B-4	CIDEr
0.8326	0.6483	0.5382	0.4919	0.8563
ROUGE Scores				
R-1	R-2	R-L	–	–
0.9422	0.9003	0.9304	–	–

Figure 8 demonstrates the image captioning and story generation system through an interactive GUI. A caption is automatically generated by the trained model once users choose an image from the dropdown menu. The caption is then tokenized and simplified. A button enables users to generate a story from the caption, assisted by GPT-4. The output is summarized and displayed, allowing for an intuitive and engaging image-to-text experience.

A. Comparative analysis and discussion

The following Table III and IV provides the comparative analysis between various models for image captioning accord-

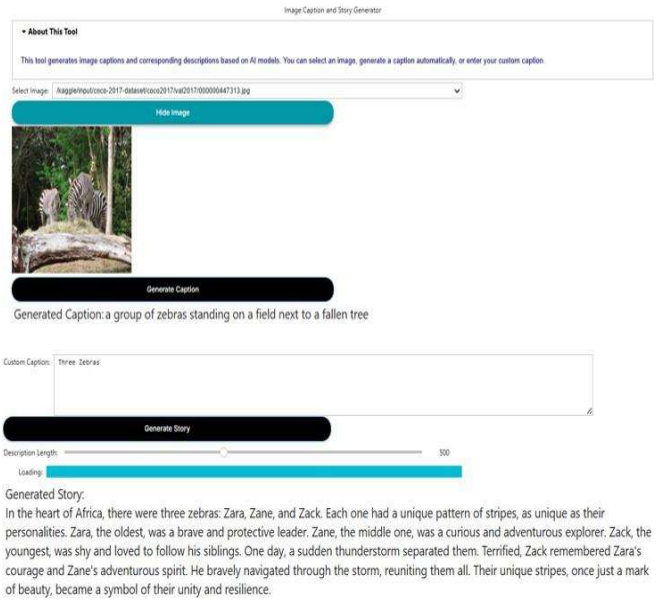


Fig. 8. Graphical User Interface for Generating Image Caption and Story

ing to BLEU, ROUGE and CIDEr performance evaluation. Table V showcases the qualitative experiments results of their method.

TABLE III
COMPARISON OF BLEU SCORES (B-1 TO B-4) ACROSS MODELS

Model	B-1	B-2	B-3	B-4
GRU [24]	0.7800	0.5700	0.4400	0.3600
EC+SI-EFO [25]	0.7666	0.5801	0.4352	0.2629
Double Attn [26]	0.8460	0.6450	0.5240	0.3620
Transformer NSC [27]	0.8070	0.6560	0.5130	0.3940
X Transformer [28]	0.8090	0.6580	0.5150	0.3970
Ens Caption [29]	0.8170	0.6530	0.5110	0.3750
PAG Net [30]	0.8320	0.6280	0.4630	0.4080
Hybrid Model (Proposed)	**0.8326**	**0.6583**	**0.5382**	**0.4919**

TABLE IV
COMPARISON OF ROUGE AND CIDER SCORES ACROSS MODELS

Model	R-1	R-2	R-L	CIDEr
GRU [24]	-	-	0.5900	1.1050
EC+SI-EFO [25]	-	-	-	-
Double Attn [26]	-	-	0.6230	133.0
Transformer NSC [27]	-	-	0.5870	129.6
X Transformer [28]	-	-	0.5910	-
Ens Caption [29]	-	-	0.5820	-
PAG Net [30]	-	-	0.5860	118.6
Hybrid Model (Proposed)	**0.9422**	**0.9003**	**0.9304**	**0.8563**

TABLE V
QUALITATIVE RESULTS OF THE PROPOSED METHOD (SINGLE-COLUMN FORMAT)

Sample Image: 
Generated Caption: a jet airplane is flying through the sky the crust
Reference Caption: a jet airplane is flying through a sky
Discussion: Reasonable match; minor wording discrepancy.
Sample Image: 
Generated Caption: a boy is out on the park flying a kite
Reference Caption: a boy is out on the park flying a kite
Discussion: Perfect match between generated and reference captions.
Sample Image: 
Generated Caption: a jeep with a deceased bird on the bathroom pass
Reference Caption: a jeep with a deceased bird on the windscreen
Discussion: Moderate alignment; discrepancy in key nouns ('bathroom' vs 'windscreen').
Sample Image: 
Generated Caption: plates loaded with some dinner and dessert with two glasses
Reference Caption: plates loaded with Thanksgiving dinner and dessert with two glasses
Discussion: Good alignment; 'Thanksgiving' replaced by 'some', reducing specificity.

V. Conclusion and Future Scope

The proposed image captioning system integrates techniques—including an Encoder-Decoder architecture, Global attention, token-level adaptive attention, multi-head attention, and EfficientNetB0 for feature extraction—and shows better performance on the COCO-2017 dataset, with high BLEU, ROUGE, and CIDEr scores. With a test accuracy of 79.87%, the model generates grammatically sound and semantically relevant captions, effectively capturing key aspects of images. Despite its success, challenges remain with complex or highly diverse image content. While the architecture surpasses many existing models on BLEU and ROUGE metrics, some alternatives achieve higher CIDEr scores, indicating room for improvement in generating more human-like captions. Additionally, the inclusion of EfficientNetB0 increases model complexity, potentially limiting real-time or low-resource applicability. Future work should aim to reduce computational overhead, enhance semantic diversity, and improve generalization. This could involve testing alternative backbone networks, refining linguistic transformations, and adopting improved multimodal fusion techniques. Expanding class coverage could also improve accuracy across a broader range of image types.

References

[1] Omri, M., Abdel-Khalek, S., Khalil, E. M., Bouslimi, J., and Joshi, G. P. 2022. Modeling of Hyperparameter Tuned Deep Learning Model for Automated Image Captioning. *Mathematics*. doi: 10.3390/math10030288.

[2] Atliha, V., and S˘es˘ok, D. 2022. "Image-Captioning Model Compression." *Appl. Sci.*, doi: 10.3390/app12031638

[3] Hossain, M. D. Z., Sohel, F., Shiratuddin, M. F., and Laga, H. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys*. doi: 10.1145/3295748.

[4] Revathi, B. S., and Meena, K. A. 2022. A review on image captioning system from artificial intelligence, machine learning and deep learning techniques. *i-manager's Journal of Image Processing*. doi: 10.26634/jip.9.3.19054.

[5] Deepak, G., Gali, S., Sonker, A., Jos, B. C., Daya Sagar, K. V., and Singh, C. 2023. Automatic image captioning system using a deep learning approach. *Soft Comput.*, doi: 10.1007/s00500-023-08544-8.

[6] Javanmardi, S., Latif, A. M., Sadeghi, M. T., Jahanbanifard, M., Bonsangue, M., and Verbeek, F. J. 2022. Caps Captioning: A Modern Image Captioning Approach Based on Improved Capsule Network. *Sensors*. doi: 10.3390/s22218376.

[7] Rinaldi, A. M., Russo, C., and Tommasino, C. 2023. Automatic image captioning combining natural language processing and deep neural networks. *Results Eng.*, doi: 10.1016/j.rineng.2023.101107

[8] Oluwasammi, A., et al. 2021. Features to text: A comprehensive survey of deep learning on semantic segmentation and image captioning. *Complexity*, doi: 10.1155/2021/5538927.

[9] Ayoub, S., Gulzar, Y., Reegu, F. A., and Turaev, S. 2022. Generating Image Captions Using Bahdanau Attention Mechanism and Transfer Learning. *Symmetry (Basel)*, doi: 10.3390/sym14122681.

[10] Yu, J., Li, J., Yu, Z., and Huang, Q. 2020. Multimodal Transformer with Multi-View Visual Representation for Image Captioning. *IEEE Trans. Circuits Syst. Video Technol.*, doi: 10.1109/TCSVT.2019.2947482.

[11] Parvin, H., Naghsh-Nilchi, A. R., and Mohammadi, H. M. 2023. Transformer-based local-global guidance for image captioning. *Expert Syst. Appl.* doi:10.1016/j.eswa.2023.119774.

[12] Wei, H., Li, Z., Zhang, C., and Ma, H. 2020. The synergy of double attention: Combine sentence-level and word-level attention for image captioning. *Comput. Vis. Image Underst.* doi:10.1016/j.cviu.2020.103068.

[13] Islam, Z., Saha, S., Islam, T., and Latif, S. 2022. Bengali Caption Generation for Images Using Deep Learning. In *Proceedings of 2022 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering, WIECON-ECE 2022*. doi:10.1109/WIECON-ECE57977.2022.10150494.

[14] Indumathi, N., Divyalakshmi, R. J., Stalin, J., Ramachandran, V., and Rajaram, P. 2023. Apply Deep Learning-based CNN and LSTM for Visual Image Caption Generator. In *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2023*. doi:10.1109/ICACITE57410.2023.10183097.

[15] Goel, N., Arora, A., Kashyap, P., and Varshney, S. 2023. An Analysis of Image Captioning Models using Deep Learning. In *2023 International Conference on Disruptive Technologies, ICDT 2023*. doi:10.1109/ICDT57929.2023.10151421.

[16] Jain, Y. S., Dhopeshwar, T., Chadha, S. K., and Pagire, V. 2021. Image Captioning using Deep Learning. In *2021 International Conference on Computational Performance Evaluation (ComPE)*, pp. 40–44. doi:10.1109/ComPE53109.2021.9751818.

[17] Rakshith, N., Gowda, M. B. K., Preetham, N., Tejas, M., and Baig, M. I. 2024. Deep Learning Hybrid Technique for Generation of Image Caption. In *2024 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT)*, pp. 1–6. doi:10.1109/IConSCEPT61884.2024.10627857.

[18] Biradar, V. G., M. G., Agarwal, S., Singh, S. K., and Bharadwaj, R. U. 2023. Leveraging Deep Learning Model for Image Caption Generation for Scenes Description. In *2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT)*, pp. 1–5. doi:10.1109/EASCT59475.2023.10393602.

[19] Sudhakar, J., Iyer, V. V., and Sharmila, S. T. 2022. Image Caption Generation using Deep Neural Networks. In *2022 International Conference for Advancement in Technology, ICONAT 2022*. doi:10.1109/ICONAT53423.2022.9726074.

[20] Kushwaha, R., and Biswas, A. 2021. Hybrid Feature and Sequence Extractor based Deep Learning Model for Image Caption Generation. In *2021 12th International Conference on Computing Communication and Networking Technologies, ICCCNT 2021*. doi:10.1109/ICCCNT51525.2021.9579897.

[21] Bansal, P., Malik, K., Kumar, S., and Singh, C. 2023. EfficientNet-based Image Captioning System. In *Proceedings of 2023*, pp. 643–647. doi:10.1109/DICCT56244.2023.10110117.

[22] Tsuchiya, G. 1971. Postmortem Angiographic Studies on the Intercoronary Arterial Anastomoses.: Report I. Studies on Intercoronary Arterial Anastomoses in Adult Human Hearts and the Influence on the Anastomoses of Strictures of the Coronary Arteries. *Jpn. Circ. J.*, vol. 34, no. 12, pp. 1213–1220. doi:10.1253/jcj.34.1213.

[23] Wang, H., Zhang, Y., and Yu, X. 2020. An overview of image caption generation methods. *Comput. Intell. Neurosci.*, vol. 2020. doi:10.1155/2020/3062706.

[24] Khan, R., Islam, M. S., Kanwal, K., Iqbal, M., Hossain, M. I., and Ye, Z. 2022. A Deep Neural Framework for Image Caption Generation Using GRU-Based Attention Mechanism. Available: http://arxiv.org/abs/2203.01594.

[25] Padate, R., Jain, A., Kalla, M., and Sharma, A. 2023. Image caption generation using a dual attention mechanism. *Eng. Appl. Artif. Intell.* doi:10.1016/j.engappai.2023.106112.

[26] Parvin, H., Naghsh-Nilchi, A. R., and Mohammadi, H. M. 2023. Image captioning using transformer-based double attention network. *Eng. Appl. Artif. Intell.* doi:10.1016/j.engappai.2023.106545.

[27] Luo, R. 2020. A better variant of self-critical sequence training. *arXiv preprint* arXiv:2003.09971.

[28] Pan, Y., Yao, T., Li, Y., and Mei, T. 2020. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10971-10980.

[29] Yang, M., et al. 2020. An Ensemble of Generation- and Retrieval-Based Image Captioning With Dual Generator Generative Adversarial Network. *IEEE Transactions on Image Processing*, 29, 9627-9640. doi:10.1109/TIP.2020.3028651.

[30] Song, L., Liu, J., Qian, B., and Chen, Y. 2019. Connecting Language to Images: A Progressive Attention-Guided Network for Simultaneous Image Captioning and Language Grounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 8885-8892. https://doi.org/10.1609/aaai.v33i01.33018885.