Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 7, 2025

A HYBRID IMAGE CAPTIONING FRAMEWORK WITH EFFICIENTNETB0 AND TRANSFORMER NETWORKS

Yasir Afnan[1], Kifayat Ullah[2], Bilal Ur Rehman[*3], Inam Ul Hassan[4], Maria Zulfiqar[5], Zawish Asif[6], Wasim Habib[7], Muhammad Amir[8], Muhammad Arshad[9]

*[1,2,*3,4,5,6,7,8]Department of Electrical Engineering, University of Engineering and Technology, Peshawar, Pakistan*
[9]Department of Industrial Engineering, University of Engineering and Technology, Peshawar, Pakistan

DOI: https://doi.org/10.5281/zenodo.16352639

Keywords

Article History
Received on 23 April 2025
Accepted on 08 July 2025
Published on 23 July 2025

Corresponding Author: *
Bilal Ur Rehman

Abstract
This research investigates an image captioning model that integrates EfficientNetB0 for feature extraction and a Transformer-based encoder-decoder for caption generation, Leveraging the balanced scaling of EfficientNetB0, the model efficiently captures rich visual representations, which are then translated into descriptive textual captions by the Transformer architecture, The proposed model is trained and evaluated on the Flickr8k dataset, utilizing 85% of the data for training, The results show that the model achieves 58% captioning accuracy after 35 epochs and attains a BLEU score of 0,75, showing competitive performance while maintaining a relatively lightweight architecture.

INTRODUCTION

Image captioning involves generating textual descriptions for images, It plays a vital role in a range of applications, including image indexing, accessibility for visually impaired users, and human-computer interaction, Traditional approaches primarily rely on convolutional neural networks (CNNs) for feature extraction and recurrent neural networks (RNNs) or long short-term memory (LSTM) networks for sentence generation [1], [2], Despite these models achieving some success, they continue to struggle with understanding complex scenes as well as multi-faceted tasks with a long time horizon, The latest improvements in transformer implementations have transformed the landscape for sequence modeling problems in NLP, offering parallel processing and the ability to capture both local and global dependencies through the self-attention mechanism [3], At the same time, more advanced CNN systems, such as EfficientNet, are available, which outperform traditional networks in accuracy while requiring fewer parameters [4], In this study, we propose an advanced image captioning system that utilizes EfficientNetB0 as a backbone for image feature extraction and employs a custom transformer encoder-decoder architecture for caption generation, The key objective here is to increase both the accuracy and the fluency of the captions produced by utilizing the advantages of both architectures,

The contributions are summarized as follows:

- EfficientNetB0 is employed as the visual feature extractor to obtain high-level semantic representations from input images,
- A customized Transformer-based encoder-decoder architecture is explicitly developed to address the image captioning problem,
- The effectiveness of the proposed model is assessed using a standard benchmark dataset, with its performance compared against existing approaches through BLEU scores and accuracy evaluation,

LITERATURE REVIEW

Significant research has been conducted in the area of image captioning, Earlier approaches primarily focused on template-based and retrieval-based methods, With deep learning gaining momentum, the focus gradually shifted to encoder-decoder architectures, which brought significant improvements in generating detailed and accurate descriptions of images, Vidyadevi G, Biradar et al, [1] present a thorough review of state-of-the-art deep learning approaches for captioning, They presented a model that combines a CNN to extract image features and an LSTM to generate a sentence, detecting objects and their relations to produce a meaningful description, This contribution revealed the potential of uniting visually grounded and language models and obtained appealing results over the Flickr8k dataset, However, the model relies on the Flickr8k data, which is not nearly large or diverse enough to serve as a robust general model, Research by Yiyu Wang et al, [2] first discussed the limitations of CNN-LSTM architectures for image captioning and then introduced a transformer-based model, The model avoided the two-stage approach by using end-to-end integration with Swin Transformer, THE MSCOCO dataset was used, The results showed considerable enhancements in CIDEr scores, Nonetheless, the model sets up a Gaussian distribution to align modalities, which can be an oversimplification of intricate feature interactions, Simao Herdade et al, [3] present the Object Relation Transformer, which enhances image captioning by capturing structural object relationships detected from images through geometric attention, This approach boosts performance on the MSCOCO dataset across the standard evaluation metrics, Nevertheless, the model utilizes spatial information only in the encoder, thereby constraining its effectiveness at decoding, Additionally, the absence of explicit word-to-object region alignment affects both interpretability and performance, He et al, [4] presented the Image Transformer, a model that generalizes the Transformer architecture for image captioning, with an emphasis on spatial relationships between image regions, Unlike conventional text-based image caption models, the Image Transformer modifies the encoding and decoding processes to be compatible with image features, The model achieves top performance on the MSCOCO dataset, demonstrating its ability to generate captions from images, Nonetheless, the model comes at the cost of high computational effort and is therefore not very suitable for real-time use, Its generalization to diverse or unseen image domains remains partially untested, Yingwei Pa et al, [5] introduced X-Linear Attention Networks (X-LAN), which use X-Linear attention blocks to capture complex interactions between visual and language features, This approach enhances image captioning by effectively combining spatial and channel-wise attention, resulting in improved performance on the COCO benchmark dataset, Nonetheless, COS-Net's dependency on CLIP for semantic retrieval can introduce training data bias, which can impact the diversity and precision of the captioned texts, Moreover, the model's performance on other datasets, except MS-COCO, has yet to be extensively tested; therefore, its generalizability to different image domains is questionable, CPTR (CaPtion TransformeR), a model that employs a complete Transformer architecture with raw images as sequential inputs, is presented by Wei Liu et al, [6], The CPTR model does not utilize convolutional layers; instead, it leverages global context from the outset, in contrast to the CNN+Transformer method, On the MSCOCO dataset, it outperforms conventional techniques and provides helpful visualizations of word and image patch interactions, However, because this method relies on the quality of external sentence corpora and visual concept detectors, its efficacy across various domains remains unknown, Cornia et al, [7] conducted an in-depth investigation into transformer-based image captioning models with a focus on their interpretability, They present analytical techniques that reveal how attention mechanisms interact with image regions during the caption generation process, Using attribution-based visualization tools, this study highlights how the model's focus shifts over time as each word is predicted, Their evaluation of different transformer variants using both CNN and Vision Transformer features offers key insights into how architectural components influence interpretability, These findings emphasize the importance of grounding in image captioning, particularly for reducing hallucinations, However, this work primarily serves as a diagnostic framework, It does not propose

solutions to directly improve model accuracy or alignment, leaving scope for future improvements in explainable vision-language systems,

Several studies have significantly advanced the field of image captioning by proposing diverse models and evaluation frameworks, Tan et al, [8] introduced a compact attention-based model designed to enhance efficiency without compromising performance, Hodosh et al, [9] framed image captioning as a ranking problem, highlighting the role of data and evaluation metrics, Tran et al, [10] emphasized rich caption generation in uncontrolled environments, while Devlin et al, [11] explored language model behaviors and effective strategies in captioning tasks, Recent developments, such as CLIP-M-Cap by Chen and Zhang [12], have integrated vision-language pretraining for improved captioning, Singh et al, [13] and Jaiswal et al, [14] have applied AI and cognitive IoT methods to enhance semantic understanding and caption quality,

The reviewed studies highlight significant improvements in image captioning, particularly with transformer-based models and attention mechanisms, However, most of these approaches focus on datasets such as MSCOCO, leaving room for testing on datasets such as Flickr8k, Additionally, many existing models rely on multi-stage processes or CNN-based feature extraction, which limits their efficiency, This study proposes a novel approach using a transformer-based architecture with custom feature extraction and an integrated attention mechanism, Unlike traditional CNN methods, our model integrates end-to-end training to enhance both visual and language feature interactions, aiming for improved efficiency and performance, particularly for datasets like Flickr8k,

METHODOLOGY

Figure 1 illustrates the flow diagram of the proposed methodology, The first step is to propose an advanced image-captioning model, envisioned as a dynamic partnership, Where CNN processes the image content, and the transformer handles the language, The efficient B0 CNN serves as the model's visual processor, scanning the images and extracting key features, These image features are then passed to the transformer, which generates coherent captions, This approach effectively addresses the major hurdles connecting visual perception and language, The CNN converts complex images into feature vectors, essentially "visual summaries," whereas the transformer excels at the relationship between these features and human-readable text,

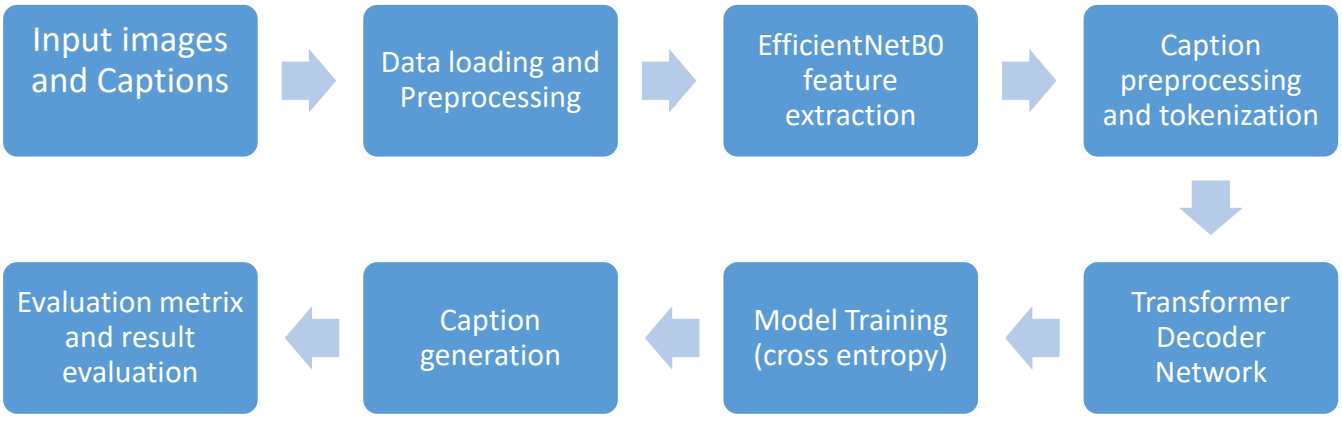

Figure 1: Flow diagram of proposed methodology

Performance Comparison EfficientNetB0

EfficientNetB0 is selected due to its superior efficiency, as clearly shown in Table 1, It achieves high accuracy with fewer parameters and faster speed, making it a better choice than models like ResNet,

Table 1. Model Performance Comparison

Model	Parameter	Accuracy	Speed
EfficientNetB0	5,3M	High	Faster
ResNet	25,6M	Similar	Slower

TRANSFORMER BASED ENCODER AND DECODER

Transformers have progressed in processing sequential data, making them an ideal choice for image captioning for the following reasons:

- Transformers handle long-range dependencies far better than older models such as RNNs or LSTM,
- They process information in parallel, which speeds up both training and inference,

1. Dataset Flicker8k:

The Flicker 8k dataset [x], which contains 8000 images in JPEG format with various shapes and sizes, and five captions each, was used, It provides a rich and diverse textual description of the image content, These captions encompass a diverse range of contexts, including people, animals, and natural scenes, making the dataset highly valuable for training and evaluating image caption models, It retains 85% of the data for training, and due to the limited size of the dataset, it uses the same subset of images for both validation and test data, Figure 2 represents a few samples of images from the dataset with captions,

Figure 2: Few samples of images from dataset with captions

2. EfficientnetB0:

EfficientNetB0, a convolutional neural network pre-trained on the ImageNet dataset, was utilized for image feature extraction due to its computational efficiency and high performance, Although the complete model comprises 237 layers, only the penultimate layer, located immediately before the final classification layer, was used, This layer outputs a 1280-dimensional feature vector that effectively captures the essential visual characteristics of each image,

During preprocessing, all images were resized to 244 × 244 pixels to meet the input size requirements of EfficientNet B0, Training was performed using a batch size of 64 to ensure stability, and the number of epochs was fixed at 30 to allow for proper convergence and optimized performance, The vocabulary size was limited to 10,000 words, and caption sequences were standardized to a maximum length of 30 tokens, Both image features and text tokens were embedded into a 512-dimensional space, with the feed-forward network dimension also set to 512 to support consistent and efficient data processing,

To enhance model robustness and reduce the risk of overfitting, image augmentation techniques such as random flipping, rotation, and contrast adjustments were applied, For textual data, captions were preprocessed by converting to lowercase, removing punctuation, tokenizing into integer sequences, and applying padding to ensure uniform length, Captions shorter than 05 tokens or longer than 30 tokens were excluded from the dataset,

This setup enabled EfficientNetB0 to extract semantically rich visual features, which were subsequently used for training the image captioning model based on a transformer architecture, as described in the following section,

3. Transformer Encoder Block:

The transformer encoder block was designed to process high-level image features and capture global contextual relationships within the data, Initially, features were extracted using the EfficientNetB0 model and subsequently passed to the encoder block, The encoder employs single-head self-attention, allowing the model to attend to different parts of the input feature sequence simultaneously, thereby enhancing its ability to model long-range dependencies and contextual information, Formally, let the input be a tensor $X \in R^{n\times d}$, where n is the sequence length and d is the embedding dimension,

The input is first normalized:

$$\hat{X} = \text{layernorm}(X) \quad (1)$$

then passed top feedforward layer with ReLU activation:

$$F = \text{ReLU}(W_1 \hat{X} + b_1) \quad (2)$$

for self attention , the same input is used s Query(Q),Key(K) AND Value (V) ,the scaled dot product is attention is computed as A

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (3)$$

Since only one attention head is used (h=1),this simplifies the multihead attention mechasnism to a single attention pathway

The result is then added to input using residual connection, followed by another normalization step:

$$Y = \text{Layer norm}(\text{Attention}(F, F, F)) \quad (4)$$

This structure allows our model for long range dependencies within the image feature sequence while maintaining computational efficiency,

4. Positional Embedding Layer:

As transformers lack an inherent sense of sequential order, which is necessary for generating captions word by word, they are not suitable for this task, So, for this Positional Embedding layer, information about token positions within the input sequence is incorporated, The layer consists of two learned embedding matrices: One for token indices and one for the respective positions,

Let x_i represent the input token at position I ,the embedding for each token is computed as:

$$E_i = \text{Embedding}(x_i), \sqrt{d} \quad (5)$$

Where d is the embedding dimension, and the multiplication with $\sqrt{d}$ ensures appropriate scaling of the embeddings, The positional embedding vector P_i is retrieved for each position i, and the final output to the model becomes:

$$Z_i = E_i + P_i \quad (6)$$

The result is a rich embedding that captures both the meaning of the word and its position in the sequence, it enables the decoder to generate syntactically ordered captions that align well with the visual content,

5. Decoder:

This block is responsible for generating textual captions on the visual features provided by the encoder, it uses a dual attention mechanism to capture dependencies within the caption sequence as well as alignment with the image context, so it receives the partially generated caption (as input) and the encoder output(images features), and predict the next word token,

The architecture of the decoder block is composed of the following components:

5.1 Self-Attention Layer (Masked):
This layer processes the decoder's own input sequence, A causal mask is applied to prevent the model from attending to future tokens during training, This ensures autoregressive generation by preserving the left-to-right ordering of words,

$$\text{Attention}_{\text{self}}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + M\right)V \quad (7)$$

where M is the causal mask and Q=K=V=Input embeddings

5.2 Cross-Attention Layer (Encoder-Decoder Attention):
After self-attention, the output attends to the encoder outputs (image features), This enables the decoder to focus on relevant parts of the image while generating each word in the caption,

$$\text{Attention}_{\text{cross}}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (8)$$

Here, Q=Decoder self-attention output, and K=V=Encoder outputs

5.3 Feed-Forward Network:
A two-layer feed-forward network (FFN) with ReLU activation refines the representation after the attention steps, This block helps in transforming and enriching the attended features:

$$\text{FFN}(x) = \text{Dense}_2(\text{ReLU}(\text{Dense}_1(x)) \quad (9)$$

5.4 Positional Embedding Layer:
Like the encoder, the decoder uses a learned positional embedding layer to inject information about word order into the model,

5.5 Normalization and Residual Connections:
Layer normalization and residual (skip) connections are applied after each attention and FFN step to stabilize training and preserve gradient flow:

5.6 Final Output Layer:
The final output is passed through a dense layer with softmax activation to predict the next word in the caption from the vocabulary,

$$y = \text{Layer Norm}(x + \text{sublayeroutput})$$

WER (10)

This decoder architecture enables the model to generate grammatically correct and semantically relevant captions by learning dependencies within the sequence and aligning them with the visual content,

RESULTS AND DISCUSSION

The model's performance is evaluated after 10, 20, and 35 epochs, The image has three original (manual) captions, as shown in Figure 5, The idea is to test our model on multiple captions, Figure 4 results after **10 epochs**, where the captions are basic and less accurate, Figures 5 and 6 present the results after **20 epochs,** indicating improved relevance and structure, After **35 epochs,** as shown in Figures 7 and 8, the captions become more fluent and semantically accurate, reflecting a significant improvement in the model, The BLEU score and model accuracy were employed as the primary evaluation metrics, The evaluation includes results for randomly selected images from the dataset, as well as external images sourced from Google, providing comprehensive insights into the model's earning behavior,

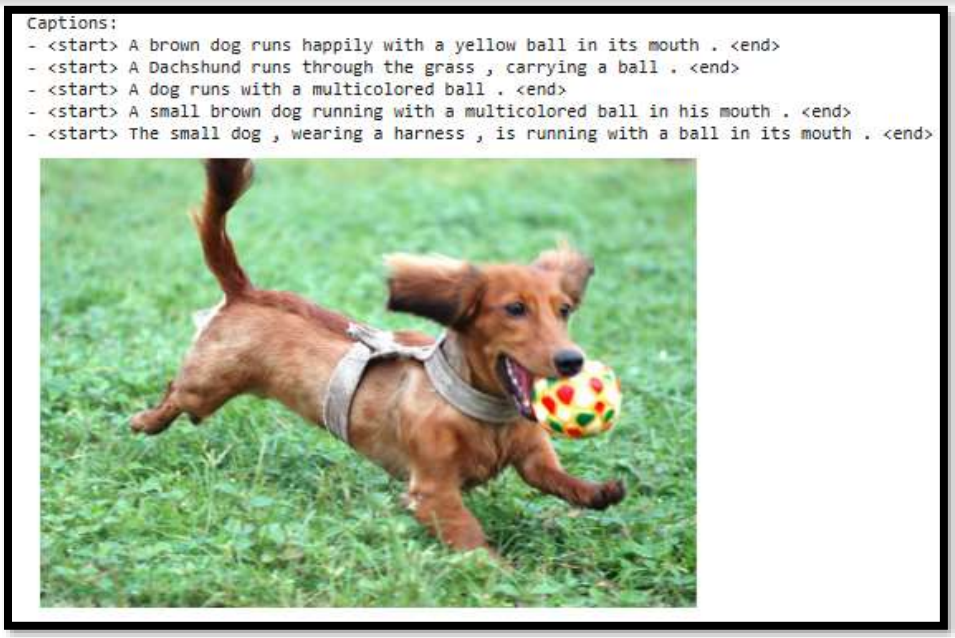

Figure 3: An image with with original captions

A-Results after 10 Epochs:

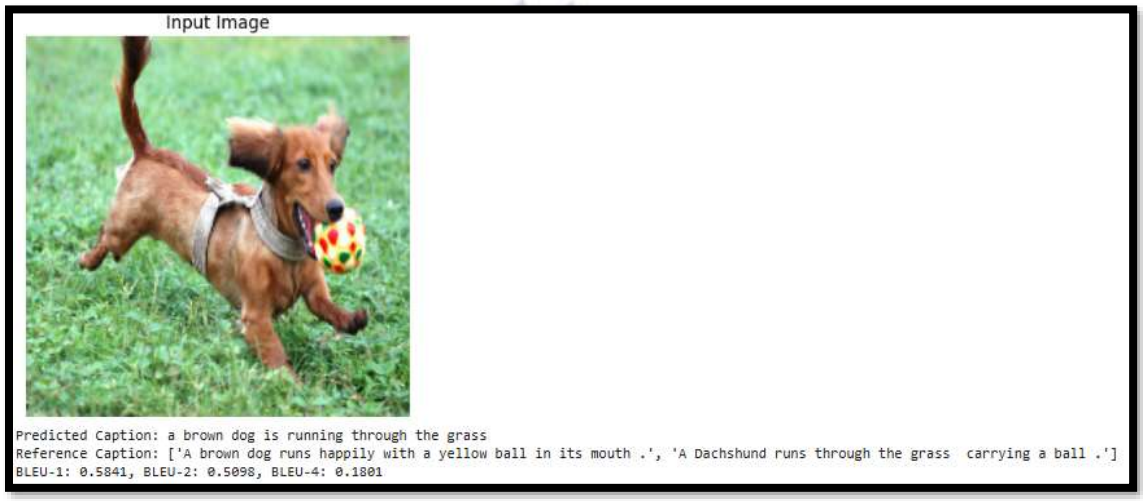

Figure 4: Image Caption Generated using Proposed Model with BLEU Scores Provided

Despite having the same accuracy and training epochs, Figure 4 achieved higher BLEU scores, indicating that the model generated a more contextually accurate and fluent caption, Hence, it highlights the model's varied performance across different image features,

The accuracy for both training and validation increases with the number of epochs, Initially low due to untrained weights, the accuracy improves as the model learns meaningful patterns from the data through gradient-based optimization,

B-Results after 20 Epochs:

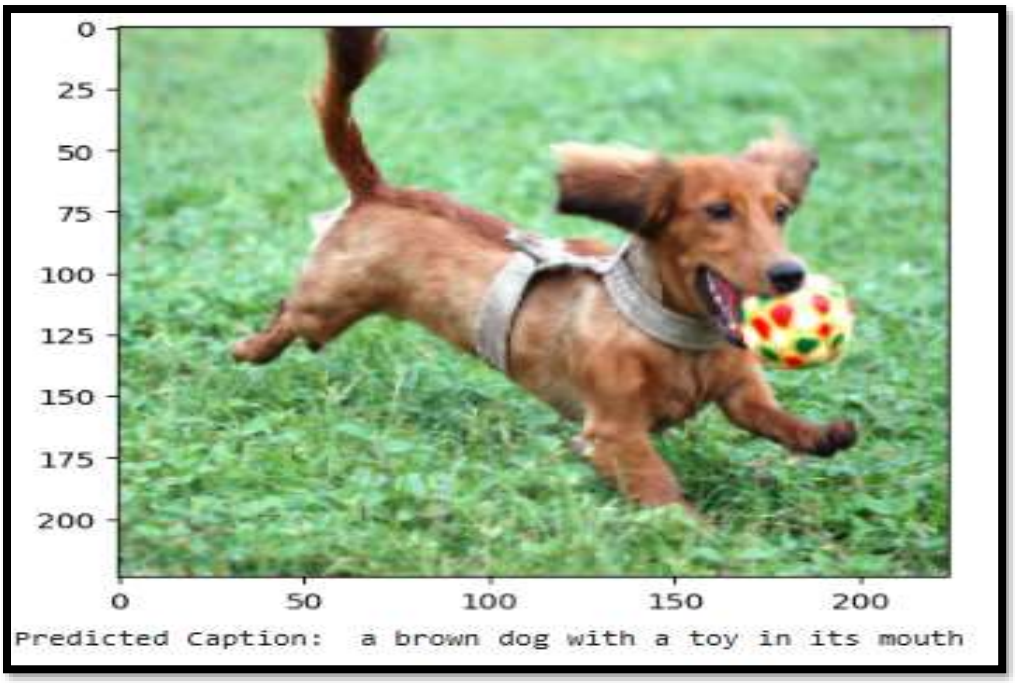

Figure 5 Caption Generated using our proposed Model

Figure 5 shows that increasing the number of epochs to 20 leads to improved caption quality and better contextual understanding, The corresponding rise in BLEU scores is presented in Figure 6, which provides the detailed evaluation metrics,

Figure 6: Caption Generated using our proposed Model with BLEU score

Based on the results shown in Figure 4, where the BLEU scores at 10 epochs were B1: 0,58, B2: 0,50, and B4: 0,18, Figure 6 presents the improved scores at 20 epochs: B1: 0,63, B2: 0,53, and B4: 0,16, The increase in BLEU-1 and BLEU-2 reflects improved word- and phrase-level accuracy, At the same time, the slight drop in BLEU-4 suggests that the model is still developing its ability to generate more extended, more

coherent captions, Hence, it reflects typical learning behavior as the model continues to optimize over more training epochs,

C-Evaluation Results at Epoch 35:

After training for 35 epochs on the Flickr8k dataset, the model achieved a training accuracy of 57%

```
Epoch 35/35
104/104 ━━━━━━━━━━━━━━━━━━━━ 79s 758ms/step - acc: 0.5760
```

Figure 7: Caption Generated using our proposed Model with 35 epochs

Figure 7 shows the same image previously used in earlier evaluation, now captioned after 35 epochs, The generated caption displays notable improvement in fluency and descriptive detail, indicating the model's enhanced ability to capture the visual context more accurately,

Figure 8: Caption Generated using our proposed Model with BLEU score

Figures 8 demonstrate a significant improvement after 35 epochs, with BLEU scores rising to B1: 0,81, B2: 0,81, and B4: 0,75, This noticeable enhancement indicates substantial progress in both word- and phrase-level accuracy, The model demonstrated an improved ability to generate more complete, fluent, and contextually appropriate captions, reflecting its

increased understanding of visual and linguistic features over time,

CONCLUSION

This research presents an image captioning system that combines EfficientNetB0 for visual feature extraction with a Transformer-based decoder to generate contextually meaningful captions, The model demonstrated progressive performance improvement, as evidenced by increasing BLEU scores and accuracy across multiple epochs, These results indicate the system's ability to learn meaningful relationships between visual input and natural language, producing captions that align well with image content, Furthermore, this research provides a strong baseline for future studies, By incorporating larger and more diverse datasets, refining the architecture, and integrating more advanced language modeling techniques, the system's accuracy, generalization ability, and caption diversity can be substantially improved,

REFERENCES

[1] V, G, Biradar, M, G, S, Agarwal, S, K, Singh and R, U, Bharadwaj, "Leveraging Deep Learning Model for Image Caption Generation for Scenes Description," 2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT), Bengaluru, India, 2023, pp, 1-5, doi: 10,1109/EASCT59475,2023,10393602,

[2] Wang, Y,, Xu, J,, & Sun, Y, (2022, June), End-to-end transformer based model for image captioning, In Proceedings of the AAAI conference on artificial intelligence (Vol, 36, No, 3, pp, 2585-2594),

[3] Huang, L,, Wang, W,, Chen, J,, & Wei, X, Y, (2019), Attention on attention for image captioning, In Proceedings of the IEEE/CVF international conference on computer vision (pp, 4634-4643),

[4] He, S,, Liao, W,, Tavakoli, H, R,, Yang, M,, Rosenhahn, B,, & Pugeault, N, (2020), Image captioning through image transformer, In Proceedings of the Asian conference on computer vision,

[5] Yao, T,, Pan, Y,, Li, Y,, Qiu, Z,, & Mei, T, (2017), Boosting image captioning with attributes, In Proceedings of the IEEE international conference on computer vision (pp, 4894-4902),

[6] Feng, Y,, Ma, L,, Liu, W,, & Luo, J, (2019), Unsupervised image captioning, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp, 4125-4134),

[7] Cornia, M,, Baraldi, L,, & Cucchiara, R, (2022), Explaining transformer-based image captioning models: An empirical analysis, AI Communications, 35(2), 111-129,

[8] Tan, J, H,, Chan, C, S,, & Chuah, J, H, (2019), Comic: Toward a compact image captioning model with attention, IEEE Transactions on Multimedia, 21(10), 2686-2696,

[9] Hodosh, M,, Young, P,, & Hockenmaier, J, (2013), Framing image description as a ranking task: Data, models and evaluation metrics, Journal of Artificial Intelligence Research, 47, 853-899,

[10] Tran, K,, He, X,, Zhang, L,, Sun, J,, Carapcea, C,, Thrasher, C,, ,,, & Sienkiewicz, C, (2016), Rich image captioning in the wild, In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp, 49-56),

[11] Devlin, J,, Cheng, H,, Fang, H,, Gupta, S,, Deng, L,, He, X,, ,,, & Mitchell, M, (2015), Language models for image captioning: The quirks and what works, arXiv preprint arXiv:1505,01809,

[12] Chen, S,, & Zhang, J, (2023, October), CLIP-M-Cap: CLIP Mean Teacher for Image Captioning, In Proceedings of the 2023 2nd International Symposium on Computing and Artificial Intelligence (pp, 50-55),

[13] Singh, Y, P,, Ahmed, S, A, L, E,, Singh, P,, Kumar, N,, & Diwakar, M, (2021, April), Image captioning using artificial intelligence, In Journal of Physics: Conference Series (Vol, 1854, No, 1, p, 012048), IOP Publishing,

[14] Jaiswal, T,, Pandey, M,, & Tripathi, P, (2021), Image captioning through cognitive IOT and machine-learning approaches, Turkish Journal of Computer and Mathematics Education, 12(9), 333-351.