Turkish Journal of Engineering

https://dergipark.org.tr/en/pub/tuje

e-ISSN 2587-1366

Exploring Bengali Image Descriptions through the combination of diverse CNN Architectures and Transformer Decoders

Biswajit Patra[*1], **Dakshina Ranjan Kisku** [1]

[1], *National Institute of Technology Durgapur, Department of Computer Science and Engineering, India, bp.21cs1102@phd.nitdgp.ac.in; drkisku.cse@nitdgp.ac.in*

Cite this study: Patra, B., & Kisku ,D,R (2025). Exploring Bengali Image Descriptions through the combination of diverse CNN Architectures and Transformer Decoders. Turkish Journal of Engineering, 9 (1), 64-78

https://doi.org/10.31127/tuje.1507442

Keywords
Image Description
CNN
Self-attention
Transformer
Bengali text-to-speech synthesis

Research Article

Received:30.06.2024
Revised:11.08.2024
Accepted:22.08.2024
Published:20.01.2025

Abstract

In recent years, there has been growing interest among researchers in the field of image captioning, which involves generating one or more descriptions for an image that closely resembles a human-generated description. Most of the existing studies in this area focus on the English language, utilizing CNN and RNN variants as encoder and decoder models, often enhanced by attention mechanisms. Despite Bengali being the fifth most-spoken native language and the seventh most widely spoken language, it has received far less attention in comparison to resource-rich languages like English. This study aims to bridge that gap by introducing a novel approach to image captioning in Bengali. By leveraging state-of-the-art Convolutional Neural Networks such as EfficientNetV2S, ConvNeXtSmall, and InceptionResNetV2 along with an improvised Transformer, the proposed system achieves both computational efficiency and the generation of accurate, contextually relevant captions. Additionally, Bengali text-to-speech synthesis is incorporated into the framework to assist visually impaired Bengali speakers in understanding their environment and visual content more effectively. The model has been evaluated using a chimeric dataset, combining Bengali descriptions from the Ban-Cap dataset with corresponding images from the Flickr 8k dataset. Utilizing EfficientNet, the proposed model attains METEOR, CIDEr, and ROUGE scores of 0.34, 0.30, and 0.40, while BLEU scores for unigram, bigram, trigram, and four-gram matching are 0.66, 0.59, 0.44 and 0.26 respectively. The study demonstrates that the proposed approach produces precise image descriptions, outperforming other state-of-the-art models in generating Bengali descriptions.

1. Introduction

Image captioning [1] involves conveying the contextual information present in an image. While an image can contain vast amounts of information, human cognitive abilities allow us to accurately describe every detail. Given the wide range of applications for image captioning, enabling computers to compete with this human-like ability to describe images is crucial. To identify individual objects, understand the relationships between them, and describe these relationships in the form of coherent captions, specialized algorithms that integrate computer vision and language processing techniques are required. For humans, extracting image elements, identifying objects, recognizing their relationships, and linking them to the overall scene as the semantic graph is relatively straightforward in storing extensive vocabulary related to these objects. However, for machines, describing the same visuals is a much more complex task. Due to its challenging nature, image captioning is a significant area of focus in artificial intelligence, where computer vision techniques [2-5] meet language modelling [6-7] mechanisms. Computer vision models, such as CNNs and vision transformers, are employed to analyze the image, while NLP models like RNNs, LSTMs, and Transformers generate the corresponding descriptions.

A significant amount of research has been conducted on image captioning, with the majority utilizing English vocabulary. However, very few studies have focused on native languages like Bengali. Bengali is spoken by over 215 million people worldwide, with 196 million native

speakers in India and Bangladesh, making it the seventh most spoken language globally [8]. Many native Bengali speakers are not proficient in English, making it crucial to develop machine-generated descriptions of image visuals in Bengali. Additionally, audio descriptions in Bengali would greatly benefit visually impaired individuals and multitasking users in Bengali-speaking regions. This would enhance communication, accessibility, and user experience across various fields such as social media, e-commerce, education, and news media for Bengali-speaking communities.

The proposed model introduces a novel approach to Bengali image captioning by integrating advanced convolutional neural network architectures, such as EfficientNetV2S, ConvNeXtSmall, and InceptionResNet V2, with a customized Transformer in an encoder-decoder framework. This model leverages the strengths of each CNN for thorough feature extraction and utilizes an improvised Transformer to generate visual descriptions in Bengali. The use of distinct image analysis techniques from different CNN architectures ensures discriminative feature representation and effective learning of image patterns. The customized Transformer enhances caption accuracy and accelerates convergence on new data by capturing complex patterns and facilitating efficient cross-modal interactions between image and text representations. Additionally, the system provides image descriptions in both audio and text formats, increasing its practical utility. The effectiveness of the proposed framework is evaluated by comparing its performance against other state-of-the-art image captioning models using the BanCap dataset. The evaluation metrics confirm that the framework is adequate for assessing caption quality, and it successfully integrates various techniques to improve image caption generation.

The key contributions are summarized as follows: (a) The framework employs diverse CNN architectures as encoders to improve feature representation. It leverages mechanisms like compound scaling, depthwise separable convolution, and multi-scale features with deep residual connections. This approach captures discriminative and complex features, enhancing diversity, including additional information in generated descriptions of the same visual, and ensuring robustness across different visual inputs. (b) An adapted Transformer in the decoder includes a cross-attention module, facilitating effective interaction between image and text representations. This model variant, with reduced layers and parameters, mitigates overfitting risks, which is particularly beneficial for smaller datasets. The attention mechanism of the improvised Transformer ensures contextually relevant word generation aligned with specific image components. (c) The proposed framework captures the entire context of image features, produces descriptions that maintain the relationship between significant components to detail the inside content properly and is more specific to the inferred image. Following the human pattern of describing the most significant component first and moving to rest for complex visuals enhances accessibility in terms of easy understanding. (d) By focusing on the Bengali language, the proposed resource-efficient solution contributes to research in low-resource language processing, serves as a foundation for further exploration in Bengali image captioning, and can be adapted for other low-resource languages such as Turkish, Finnish, Slovenian, Hindi, Urdu, etc., subject to compatibility with the standardize English captions of Benchmark datasets.

2. Literature Survey

Early approaches for describing the image description often relied on template-based and retrieval-based methods discussed in [9-11]. In the template-based approach, after detecting the scenes, associated objects, and relationships among them, it is entered into its most convenient place in the predefined templates to generate the description. In retrieval-based approaches, the main focus is on retrieving images. This technique entails retrieving similar images and utilizing the captions associated with these retrieved images as captions for the query image. Both template-based and retrieval-based methods often suffer from a lack of generalization when generating captions. Template-based methods have limited capacity to understand the contextual nuances of an image since they rely on predefined templates, which can result in captions of restricted length. Similarly, the retrieval-based method exhibits better performance when a large corpus dataset of similar images is available. Thus, both approaches struggle to grasp the context of an image beyond similar images in the retrieval pool, limiting the diversity and originality of generated descriptions. To address these challenges, deep learning has been employed in the encoder-decoder framework to describe the visuals. Early image captioning deep model [1] often relied on a pre-trained CNN in the encoder to extract image features, which were then fed into the decoder, including RNN or its variants, to generate captions. Further, attention-based approaches [12-13] have been incorporated to refine the English description in [14-15] by focusing on the key components of the visuals. To enhance the performance of generating content-specific descriptions, the transformer [16] is employed in the succeeding work. Instead of convolutional and recurrent neural networks, the multimodal transformer [10] fully relies on the attention mechanism to assess the global dependencies between image and text representations. To enhance the performance of image descriptions, a local and global generator model based on self-attention is proposed in [11]. The Meshed-Memory Transformer model discussed in [17] learns a multi-level representation of relationships between image regions by integrating learned knowledge and utilizing low and high-level features in decoding. To generate high-quality descriptions, generative transformer-based language models like GPT-2 or GPT-3 are employed. Having an image analysis component like a vision transformer embedded in the GPT-2 model in [18-19] facilitates image caption generation efficiently. All the entities obtained from the visuals are utilized by the GPT-3-based model [20] to produce more specific descriptions. To enhance the experience of accessibility of the visuals for the blind population, an OCR-integrated reading device [21] based on deep models is introduced. A portable

device, 'Orcam MyEye' and a mobile application, 'Seeing AI', designed to provide more interactive and enhanced learning experiences for the blind population are discussed in [22]. 'Orcam MyEye' use smart magnifier glass to read from any book or screen on behalf of visually impaired students. It also recognize faces, identify products accurately and provide quick information on interaction. 'Seeing AI' narrates the world around the blind and low vision community by harnessing the power of AI to pursue daily tasks from reading to describing photos. A powerful tool, 'AR4VI' [23], was developed for visually impaired users with the potential to remove or significantly reduce a range of accessibility barriers. It imposes augmented reality to provide real-time information through audio and visual cues, helping users navigate and understand their surroundings better. Research work discussed in [21-23] focuses on studies and technologies that make generated texts more accessible to visually impaired users but not optimized for smaller datasets or languages with fewer resources. Automatic alt-text (AAT) [24] imposes computer vision to identify faces, objects, and themes from an image to generate photo alt-text for screen reader users on Facebook. Thus, it enhances accessibility for blind users on social media platforms but lacks proper interpretation of complex scenes, leading to vague or overly general captions. In [25], the VizWiz dataset is introduced, including visual questions from blind users and corresponding answers. Thus, it provides a good foundation for more user-friendly and diverse systems tailored to the specific needs of blind users and highlights the potential of visual question-answering, which can further be integrated into wearable smart reading devices, autonomous vehicles, robots, etc. In [26], accessible image descriptions enhance art engagement for people with blindness and low vision. It explores methods for creating detailed and meaningful descriptions that allow these individuals to form mental representations of artwork but face challenges in conveying visual nuances. A transformer-based framework was designed in [27] to evaluate the accessibility of image descriptions in a structured approach based on compliance with the guidelines of the National Center for Accessible Media. A dense image captioning model proposed in [28] is based on object detection and localization and generates dense descriptions with the help of LSTM. This improves the coherence and quality of image captions by increasing the model complexity while enhancing the accessibility of generated captions in the Hindi language. Since RNN or its variants are sequential in nature, the problem of sequential dependency is resolved by employing a transformer model in the decoder for large datasets to generate Hindi captions, as done in [29]. While image captioning has been extensively studied for resource-rich languages like English, relatively fewer works have targeted languages with fewer available resources, such as Hindi, Bengali, etc. In [30-31], image features are extracted using VGG16 and InceptionV3, and in consequence, the LSTM model is used to generate the description in the Bengali language. Bidirectional GRU [32-33] and Bidirectional LSTM [33], along with deep CNN, are used to describe complex scenes by leveraging information from both past and future timestamps. Further visual attention is incorporated in combination with CNN and GRU in [34] to generate context-aware Bengali descriptions of the visuals. To enrich the Bengali descriptions of the visuals with more accuracy and parallelism, a transformer model has been proposed [35]. However, it lacks effectiveness in capturing complex patterns and incorporating many layers makes the model more complex too.

3. Method

3.1. Motivation

Bengali is widely spoken in South Asia, particularly in India and Bangladesh, creating a need for an efficient image captioning model in this language. Although some deep learning models have been developed for generating Bengali descriptions in [30-33], several challenges remain. Initial models often lack context awareness, leading to less coherent and relevant descriptions. Sequential execution in these models can be time-consuming and computationally inefficient, while traditional attention mechanism-based models [34] increase context relevancy but limit context windows, restricting their ability to capture global context from images. Transformers [16], which use self-attention mechanisms [13], offer a global view of image features and more effective context modelling, but their complexity and overfitting risks, especially with small datasets, pose challenges in [10-11, 17, 29, 35]. Additionally, models like GPT, designed for text processing, require integration with visual components to generate captions [18-20], demanding significant computational resources. The need for less complex, more efficient models that can adjust their activation patterns, capture complex patterns based on the specific characteristics of image data and provide detailed descriptions from different perceptions is evident. Moreover, incorporating Bengali audio descriptions could enhance accessibility for visually impaired users and multitaskers in Bengali-speaking regions. Balancing model generalization and complexity is essential for improving efficiency and effectiveness in Bengali image captioning.

3.2. Baseline Architectures

3.2.1. Convolutional Neural Network Architectures

- **EfficientNetV2S:** This convolutional neural network [36] architecture provides a good balance between model size and performance while being computationally efficient. It attains the intended efficiency by employing a compound scaling technique, which uniformly adjusts the network's resolution, depth, and width using a predefined set of scaling coefficients.

- **ConvNeXtSmall:** It captures complex patterns in images effectively by employing a combination of grouped convolutions and cross-channel interactions as given in [37]. It

uses many parallel branches in its convolution layers. By increasing the number of parallel branches it can efficiently increase model capacity. Thus feature representation and learning capacity get enhanced without significantly increasing computational cost.

- **InceptionResNetV2:** It includes residual connections like ResNet, which aid in training deeper networks and mitigating the vanishing gradient problem. Additionally, it incorporates multi-scale feature extraction through the use of inception modules, which consist of parallel convolutional layers with different kernel sizes. Details of the Inception-Resnet architecture can be found in [38].

3.2.2. Transformer Architecture

The Transformer [16] is a deep learning architecture that leverages self-attention [13] mechanisms to assign weights to different segments of input data. As depicted in "Figure 1", it consists of N encoder and decoder blocks stacked together. An encoder block includes a multi-head self-attention block

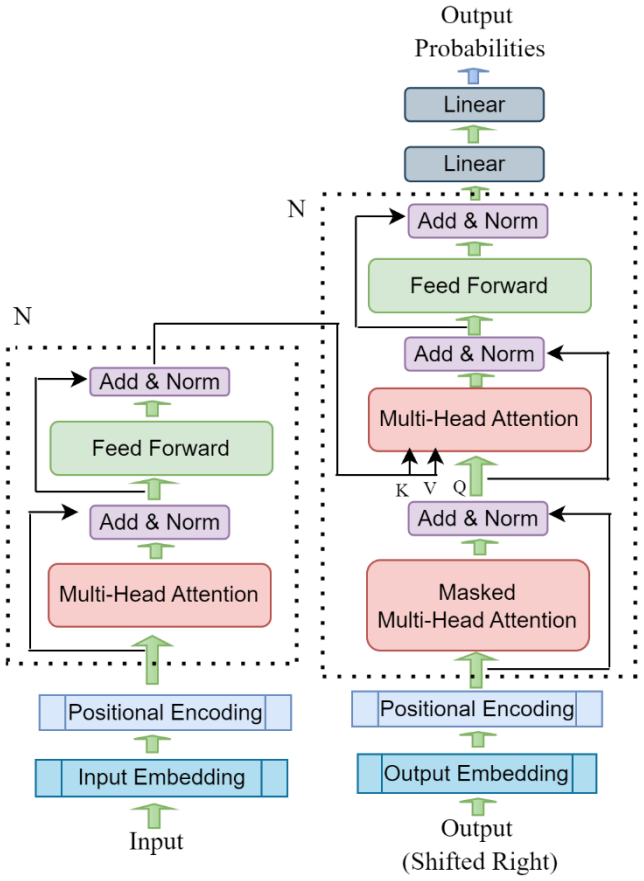

Figure 1. Transformer Architecture.

and a point-wise feed-forward network, while the transformer decoder adds one masked multi-head self-attention block to it. While the multi-head attention block consists of a padding mask, padding as well as a look-ahead mask reside in the masked multi-head attention block. Masking prevents the transformer from interpreting padding as part of the input data. Once the input is obtained in the transformer encoder, it passes through a dense layer to match it with the embedding dimensions of the transformer decoder. Further, they are merged with positional encoding and proceed through N encoder blocks. Similar to the encoder, every transformer decoder is equipped with embedding and positional encoding operations. The obtained K (key) and V (value) from the encoder are transmitted to the multi-head attention block of the transformer decoder. It receives Q (query) from its masked multi-head attention block. Attention weight is computed in its head by using query and key and further used to detect the next sequence as depicted in [16]. Multiple queries, keys, and values are obtained concurrently in its multiple heads. Since each head learns attention weights independently, every head is expected to capture distinct elements of the input sequence. Further, multiple outputs generated across multiple heads are concatenated and linearly transformed into the feed-forward block. Finally, the decoder's output is transmitted to the consecutive linear layer to obtain the probability distribution. The outputs of the feed-forward and self-attention layers are both subjected to layer normalization and residual connections in order to improve stability during training.

3.3. Proposed methodology

The proposed methodology illustrated in "Figure 2", is designed to efficiently create a contextually relevant Bengali description of images in text and audio. This framework comprises data acquisition and its processing, image encoding, linguistic decoding, and text-to-speech conversion processes, as detailed below.

3.3.1. Data Preparation

To assess and train the proposed hybrid model, an image set with its associated Bengali descriptions is required. Since no such standard Bengali captioning dataset with broader coverage of various image samples of common objects along with its Bengali description is available, two datasets are utilized in combined form. The first is the image dataset Flickr8k [1], comprising 8091 images along with five human annotated descriptions for each in English language, and the second is the Bengali caption dataset from BAN-Cap [39], consisting of corresponding Bengali translated descriptions of those 40455 English descriptions along with the image ids are used. Thus, images, along with their Bengali description, are formed from the combination of the two mentioned datasets. Flickr consists of a wide range of real-world scenes and activities, and every visual is annotated with five different captions, providing a wide variety of descriptions for the same visual content. Thus, training models using this dataset show improved performance in common object scenarios. Based on the number of provided descriptions associated with each image, multiple copies of these images are generated and randomly shuffled. Random shuffling of images along with their corresponding captions from the training dataset for each training iteration ensures that the

proposed model is exposed to a wide variety of images present in the dataset rather than biased towards learning patterns from a specific subset. Randomizing the order breaks such patterns and leading to accurate and unbiased caption generation by preventing the proposed model from following the same training sequence in subsequent iterations. This avoids the dominance of certain types of images or classes rather helps in covering different contexts, objects, and scenes and each class of images gets an equal opportunity to influence the model during training. All available special characters, punctuations, numeric values, unwanted space, the ambiguity of words, etc. are eliminated by data cleaning. Start and end tokens are added to each description as markers. Further unique words are filtered and according to the frequency of their occurrence, a vocabulary of informative words is created. Finally, word embedding of informative vocabulary is done.

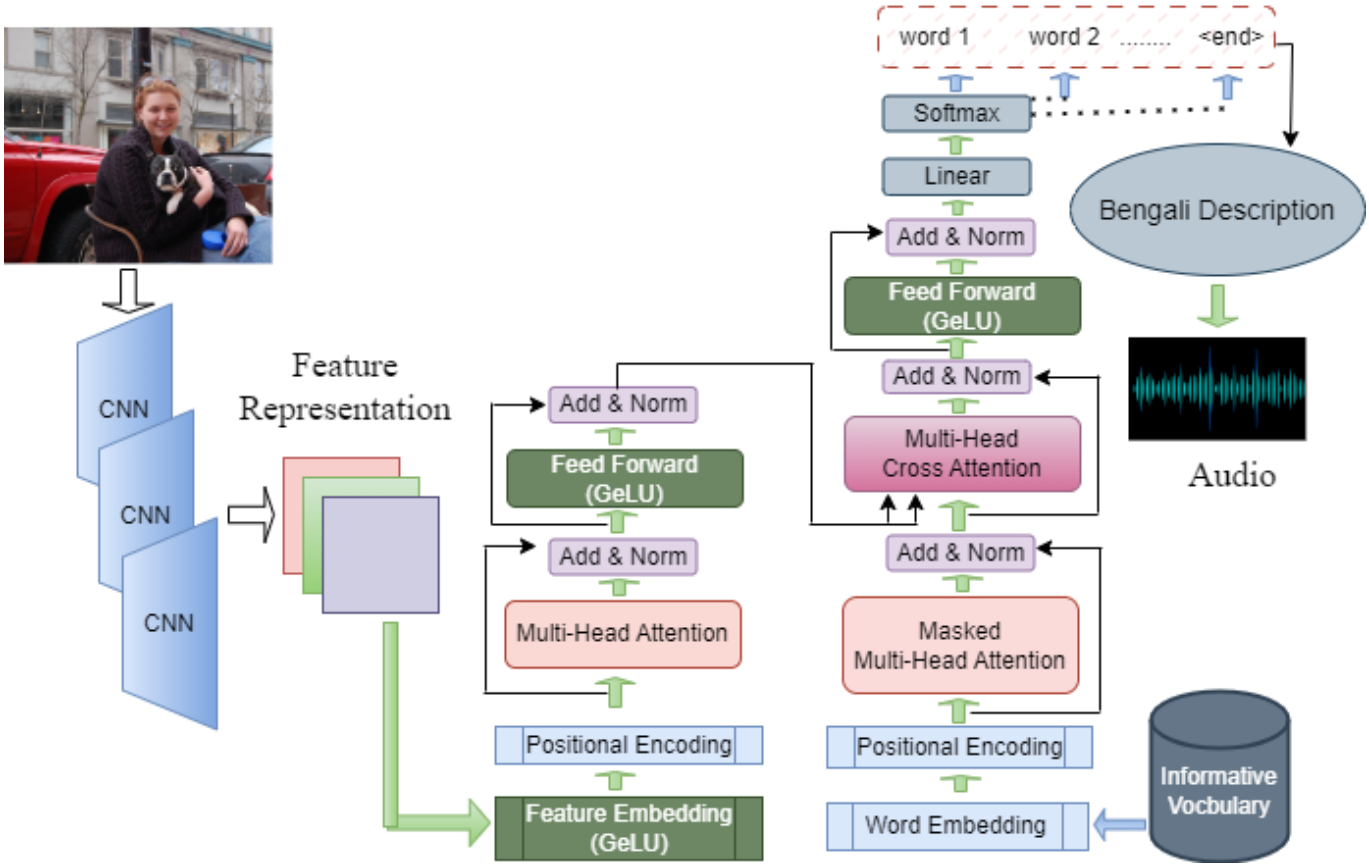

Figure 2. Proposed Bengali Description Generation Methodology.

3.3.2. Image Encoding

The proposed model can capture a wide range of image features, ensuring robustness and adaptability to various visual inputs by utilizing a diverse set of CNN architectures. The encoding process utilizes the convolutional base of the EfficientNetV2S [36], ConvNextSmall [37], and InceptionResNetV2 [38] to extract visual features from the image. The final pooling layer of the InceptionResNetV2 provides a feature map of 64 regions with each of dimension 1536, whereas, from EfficientNetV2S and ConvNextSmall, 49 image regions are obtained with each of 1280 and 768 dimensions, respectively. Feature extraction using EfficientNet leverages a compound scaling mechanism to balance the resolution, width, and depth of the network to extract rich visual features with minimal computational resources. Extracting features using ConvNext entails passing images through its depth and point-wise convolutional layers, where each convolutional operation progressively refines the input data, resulting in informative feature representations. Since Inception-ResNet uses the combination of multi-scale convolutions with residual connections, it passes images through its interconnected blocks, where features are extracted at different resolutions and levels of abstraction, facilitating robust representation learning. Further, the obtained feature representation is subjected to an embedding process [16], and it leads to an embedded feature vector, which is subsequently passed to the transformer-based decoder for post-processing.

3.3.3. Linguistic Decoding

A Transformer model with some modification is adopted in order to process the feature representation further. The proposed hybrid model uses a single encoder and decoder block inside the transformer decoder. The input embedding layer is modified into the

feature embedding layer for embedding the obtained feature representation to 512. The embedded feature map is then used as tokens for further processing. In consequence, positional encoding is performed to incorporate positional information regarding the spatial relationships among those tokens are taken as input to the encoder layer. Employing Gaussian Error Linear Unit [40] in the feature embedding layer and feed-forward blocks leads to improved training stability, better generalization, and potentially higher model performance. Self-attention is considered as the building block of the Transformer and encoder block comprises a multi-head self-attention layer and a feed-forward neural network. If X_n represents the n number of tokens, the keys (K), queries (Q), and values (V) are calculated by multiplying learnable weight vectors i.e.; W_K, W_Q, W_V with the input tokens in every attention head as given in Equation (1).

$$\begin{aligned} Q &= X_n * W_Q \\ K &= X_n * W_K \\ V &= X_n * W_V \end{aligned} \quad (\mathbf{1})$$

Attention weights determine how much each token attends to other tokens in the input sequence. The attention weights, A_w for the appearance features are then computed in the following Equation (2), by utilizing the value of the query and key from Equation (1), and a constant scaling factor, $d_k = 64$ is chosen as their dimension as the square root of the dimension of the key ensures a more stable gradient to the model.

$$A_w = Q * \frac{K^T}{\sqrt{d_k}} \quad (\mathbf{2}).$$

To make the proposed model stable during training, the attention weights across different tokens are normalized using softmax. The output of a head is computed in Equation (3), by applying the softmax function determined on the obtained attention weights, A_w from Equation (2), and further multiplied with the obtained value from Equation (1).

$$head(X) = selfattn(Q, K, V) = softmax(A_w) * V \quad (\mathbf{3})$$

The output of all heads is then concatenated and multiplied with a learned projection matrix, W_O to obtain one output vector in Equation (4) as follows:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W_O \ (\mathbf{4})$$

The Add and Normalize layer is added in conjunction with a multi-head attention block in order to control the flow of residual connections and stabilize the process of training. The next component of the encoder layer is the point-wise feed-forward net work which is applied to each output of the attention layer. It uses the GELU activation function to conduct two linear transformations by utilizing the associated weights i,e; w_1, w_2 and bias i,e; b_1, b_2 as given in Equation (5).

$$FFN(x) = GeLU(x * w_1 + b_1)w_2 + b_2 \quad (\mathbf{5})$$

The output from the fully connected layer of the encoder block is passed to the subsequent decoder of the transformer. The input embedding layer of the transformer decoder is modified to the word embedding layer and subsequent addition of positional encoding is done to process the informative vocabulary of Bengali words in appropriate order. The components of the transformer decoder remain the same as its encoder except for one extra masked attention layer to receive input words and compute as expressed in Equation (6). Similar to multi-head attention, the embedded token is analyzed, except the information needed to predict the token in the subsequent location is concealed in a matrix made up of mask values, M for "0" and "−∞".

$$Maskedattn(Q, K, V) = softmax(\frac{Q * K^T + M}{\sqrt{d_k}}) * V \quad (\mathbf{6})$$

The output from the last encoder, as well as output from the masked multi-head attention layer of the decoder, are passed to the multi-head cross-attention layer. All operations in the cross-attention layer are found to be the same as that of the multi-head attention layer, but with different modalities, inputs of image and text representations are used to mix the embedding sequences of it. The text generation process attended to different parts of the image, after getting the relevant image region dynamically for generating subsequent words. Additionally, it captures various aspects of the data in parallel, which provides faster training times, improved handling of long-range dependencies and better hardware utilization. In the Cross-attention layer, the embedded sequences of images are utilized as value (V), and key (K), and embedded sequences of text are used as query (Q) to obtain the subsequent token in the output sequence as done in [41]. Thus the cross-attention layer improves the alignment and integration of information from both modalities, leading to more accurate and contextually appropriate caption tokens. Hence, the output of the final decoder's probability distribution is used to generate the contiguous Bengali word. These steps are repeated for every new token to form the Bengali sentence.

3.3.4. Bengali text-to-speech

To generate audio descriptions along with textual descriptions in Bengali, the Bengali text-to-speech synthesis model reported in [42] is integrated with the proposed framework. It provides the audio description from the extracted linguistic features to assist visually impaired people of native regions and offers more flexibility to multitasking users. The Bengali description undergoes a few pre-processing steps to handle linguistic features such as word segmentation, punctuation, and special characters. The processed text is analyzed to identify linguistic features such as parts of speech, grammatical structure, and syntactic elements. The appropriate pitch, duration, and emphasis for each word and phrase in the text are determined by the intonation, rhythm, and stress patterns in speech. Each Bengali character or combination of characters is mapped to its corresponding phoneme representation. It ensures the

generated speech is natural and intelligible by understanding the phonological rules of Bengali, implementing a robust mapping system, and handling complex linguistic features. Acoustic models play a key role in good-quality text-to-speech solutions and are responsible for converting phoneme sequences into audible speech. It is trained using large datasets of Bengali speech to learn the relationship between phonetic units and their acoustic properties to produce natural-sounding speech from the phonetic representation of the input text. The synthesized speech waveform is generated by combining the phonetic representations of the input text with the acoustic models and further processed to enrich the quality and naturalness of the generated speech descriptions.

4. Evaluation

4.1. Experimental Setting

The proposed hybrid model of Bengali description generation of visuals is developed in Python version 3.10.12, employing the Keras framework alongside Tensorflow of version 2.15.0. Google Colab Pro facilitating NVIDIA Tesla T4 GPU with high GPU RAM is used to train the proposed hybrid model. The proposed model uses a sparse categorical cross entropy [43] to compute the difference in loss between the true class labels and the predicted class probabilities. If L is the length of the predicted description, V is the size of the informative vocabulary, $O_{t,i}$ is the true class word and $O'_{t,i}$ is the predicted word at time step t, belonging to class i, the mathematical expression is given in Equation (7) compute the cross-entropy loss as shown below.

$$L(O, O') = -1/L \sum_{t=1}^{L} \sum_{i}^{V} O_{t,i} \cdot \log(O'_{t,i}) \qquad (7)$$

The proposed model is trained with three distinct CNNs for 60 epochs, taking into account the cross-entropy loss. The dataset is divided into disjoint subsets at a ratio of 80:20 for training and validation purposes. The ratio of longest to shortest sentence in the processed captions is found to be 34:4. To maintain uniformity, all shortest-length sentences are zero-padded to make it to 34. A single encoder and decoder with GELU activation function is used in the transformer-based decoder. In image captioning tasks, the relationship between visual features extracted by the CNN and linguistic features processed by the Transformer can be complex and multifaceted. Dynamically adjusting the learning rate can help the model navigate these complexities more effectively, potentially leading to better alignment between image content and generated captions. It also ensures stability in training and speeds up the convergence, as, during the initial training phase, a higher learning rate can allow the model to quickly traverse the loss landscape and find a reasonable set of parameters to reduce the training loss. As training progresses and the landscape narrower around a local minimum, reducing the learning rate can help the proposed model converge more precisely toward an optimal solution. By annealing the learning using a custom LR scheduler based on certain conditions, the model can potentially avoid overfitting and generalize better to unseen data. Since the learning rate keeps adjusting dynamically, AdamW [44] becomes a good choice to optimize the training procedure as it is found to be less sensitive to the changing learning rate and offers better stability during training. The update rule for AdamW combines the concepts of Adam optimization with weight decay. If θ_t is the weight at time step t, λ is the weight decay coefficient, LR is the changing learning rate, β1 and β2 are the estimated biases in the first and second moment at time step t, and c is a constant, then weight at the next timestamp, θ_{t+1} can be updated as given in Equation (8).

$$\theta_{t+1} = \theta_t - LR/(\sqrt{\beta 1} + c) * \beta 2 - \lambda * LR * \theta_t \qquad (8)$$

The hyperparameters which are used in the proposed model while training are summarized in "Table 1".

Table 1. Summarization of Hyperparameters.

Hyperparameter	Batch Size	Buffer size	Embedding dimension	Dropout rate	Learning Rate (LR)	Activation function	Optimizer	Loss function
Value	64	1000	512	0.001	Custom LR scheduler	GELU	AdamW	Sparse Categorical Cross Entropy

5. Experimental Results

The proposed methodology produces contextually relevant and distinct Bengali descriptions of the visuals in an efficient manner by utilizing the combination of several SOTA CNN models and an improvised transformer in the encoder-decoder framework. The training and validation loss plot for the proposed model using distinct CNN is displayed in "Figure 3". To exhibit the performance of the proposed model built on diverse CNN and transformer decoders, Bengali descriptions generated for different images are shown in "Figure 4 (a) and 4 (b)". Moreover, on every individual image, three coherent image descriptions are produced following the proposed models. Following the generation of the Bengali description, the pivotal step involves finding the degree of similarity between the human-annotated descriptions and the model-generated captions. To validate the accuracy of the proposed models, the performance metrics, such as BLEU-n, METEOR, ROUGHE, and CIDEr are used. Among such metrics, BLEU-n measures [45] the precision of n contiguous words in the sequences of the model-generated descriptions in comparison to the ground truth descriptions and combines them using a geometric mean. METEOR [46] considers both unigram precision and recall, along with matches based on stemming and synonymy. It also incorporates penalties for unigram and word order differences and evaluates the quality of the generated sentence based on the computed F-score.

ROUGHE [47] computes recall, precision, and F1-score across different n-gram lengths and presents an overall score to assess the correspondence between the predicted and actual descriptions. Consensus-based similarity and n-gram similarity are computed to obtain the CIDEr [48] score which reflects the variation and

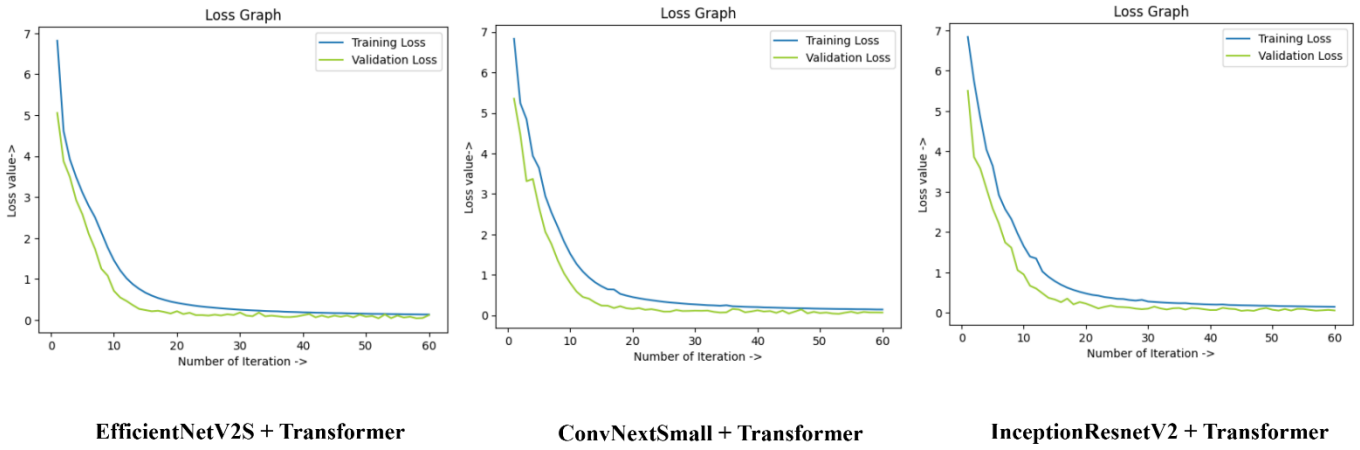

Figure 3. Loss Graph of Proposed Model using three distinct CNN.

relevance of the generated description. The assessment of the proposed models and comparisons with other SOTA models are shown in "Table 2", where BLEU-n, METEOR, ROUGHE, and CIDEr are represented as B-n, M, R, C. In the mentioned table, BLEU scores for the precision of unigram, bigram, trigram, and four-gram matches between the proposed model-generated description and the provided ground truth description are denoted by B-1, B-2, B-3, and B-4.

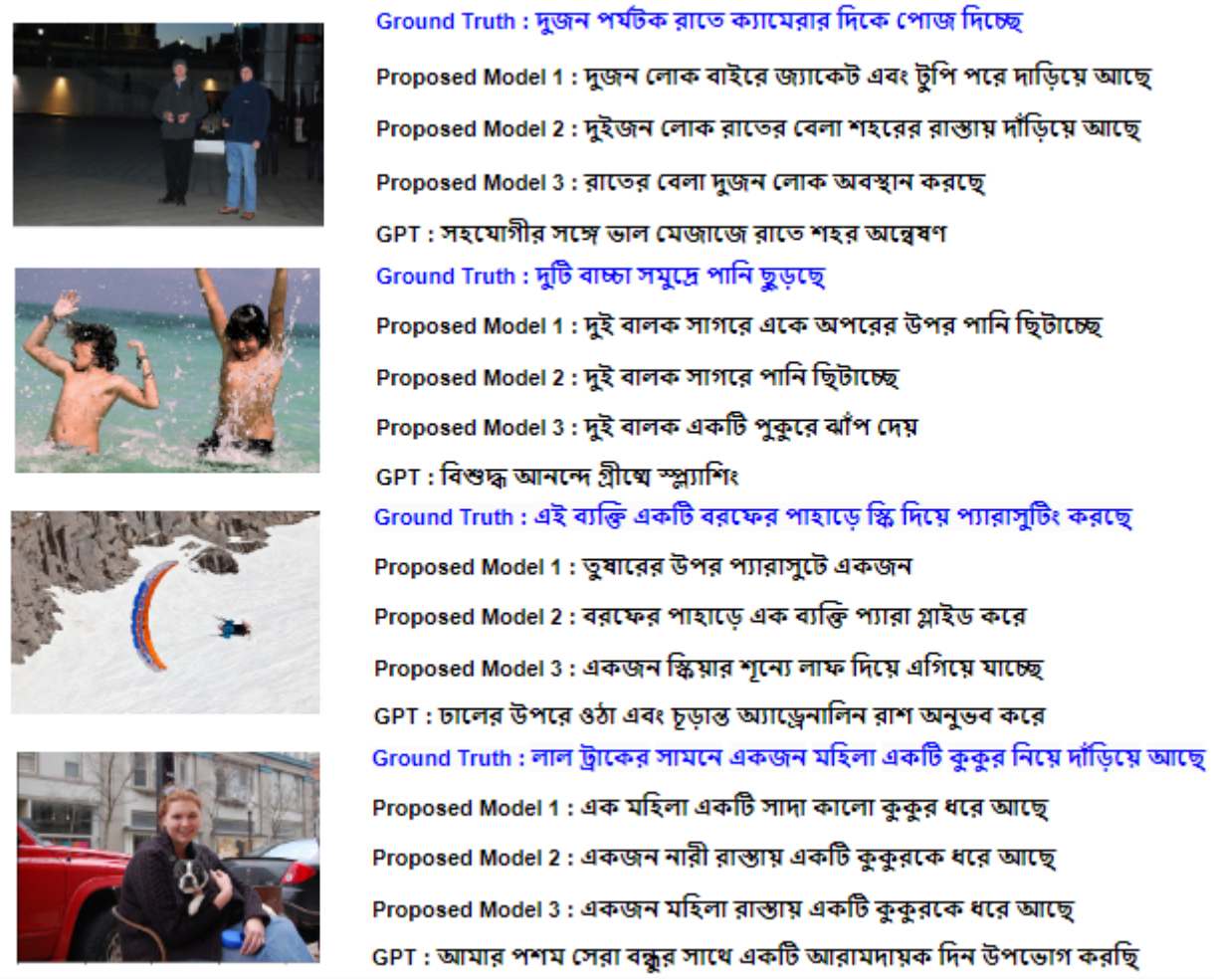

Figure 4(a). Visualization of Ground truth and Predicted Bengali descriptions.

Ground Truth : পাথুরে সৈকতে বালুতে দুজন লোক লিখছেন

Proposed Model 1 : দুই ছেলে পাথুরে সমুদ্রের ধারে বালিতে লিখছে

Proposed Model 2 : সমুদ্রের তীরে দুই ছেলে খালি গায়ে বালিতে লিখছে

Proposed Model 3 : ছেলেগুলো লাঠি দিয়ে সৈকতের বালিতে লিখছে

GPT : সৈকত দিন এবং বন্ধুত্বপূর্ণ প্রতিযোগিতা সেরা স্মৃতি তৈরি করে

Ground Truth : তিনটি কুকুর সমুদ্রে প্রবেশ করছে

Proposed Model 1 : সমুদ্রে নিক্ষিপ্ত একটি বলের জন্য তিনটি কুকুর দৌড়াচ্ছে

Proposed Model 2 : একদল কুকুর সাগরে নিক্ষিপ্ত একটি বলের জন্য দৌড়াচ্ছে

Proposed Model 3 : তিনটি কুকুর সমুদ্র সৈকতে খেলছে

GPT : কুকুরগুলো সমুদ্রে ঢেউ তাড়া করছিল

Ground Truth : অল্প বয়সী মেয়ে একটি স্লাইডের নীচে

Proposed Model 1 : সাদা জামা পরা একটি ছোট মেয়ে একটি স্লাইডে নিচের দিকে যাচ্ছে

Proposed Model 2 : একটি মেয়ে খেলার মাঠে একটি নীল স্লাইডে চড়ছে

Proposed Model 3 : একটি খেলার মাঠে নীল স্লাইডে একটি ছোট্ট মেয়ে

GPT : গ্রীষ্মের মজার মধ্যে সহচরী

Ground Truth : সাদা জামা ও চামড়ার কোট পরা এক লোক

Proposed Model 1 : সাদা জামা ও লেদার জ্যাকেট পড়া কৃষ্ণ বর্ণের লোকটি নিচে তাকাচ্ছে

Proposed Model 2 : রোদচশমা পরা কৃষ্ণ বর্ণের লোকটি নিচে তাকাচ্ছে

Proposed Model 3 : রোদচশমা ও লেদার জ্যাকেট পরা এক ব্যক্তি নিচের দিকে তাকাচ্ছে

GPT : শহরের স্পন্দন উপভোগ করা এবং রাস্তার শৈলীতে ভিজানো

Figure 4(b). Visualization of Ground truth and Predicted Bengali descriptions.

Table 2. Assessment of Proposed Model and Comparative Analysis with other SOTA models on Ban-Cap Dataset.

Methodology	B-1	B-2	B-3	B-4	M	R	C
Chittron [30]	0.56	0.35	0.22	0.13	0.28	0.29	0.17
Oboyob [31]	0.55	0.38	0.27	0.15	0.32	0.51	0.22
Hybridized using BiGRU [33]	0.53	0.35	0.24	0.12	-	-	-
Hybridized using BiLSTM [33]	0.54	0.37	0.26	0.14	-	-	-
Visual Attention Based [34]	0.58	0.36	0.25	0.14	0.29	0.28	0.19
Transformer Based [35]	0.64	0.58	0.43	0.24	0.31	0.39	0.28
Proposed Model (EfficientNetV2S)	**0.66**	**0.59**	**0.44**	**0.26**	**0.34**	**0.40**	**0.30**
Proposed Model (ConvNeXtSmall)	**0.65**	**0.58**	**0.43**	**0.25**	**0.32**	**0.39**	**0.29**
Proposed Model (InceptionResNetV2)	**0.64**	**0.56**	**0.42**	**0.24**	**0.31**	**0.39**	**0.28**

6. Discussion

The effectiveness of the image captioning model relies significantly on both the contextual information within the image and its associated captions provided in the training data. Thus integration of the right combination of deep models is crucial for generating coherent and context-specific descriptions of the visuals. İn [30], VGG-16 along with LSTM are used in the encoder-decoder framework to produce Bengali descriptions. Inception and VGG are used in [31] following different architecture. In an attempt to improve performance by capturing both the past and future context of each word in the sequence Bi-directional GRU [32-33] and Bi-directional LSTM [33] are used. To obtain a more precise description, visual attention is applied along with deep CNN and GRU [34]. "Table 2", depicts that the proposed hybrid model not only surpasses the CNN-RNN [30-33] and the attention-based [34] Bengali captioning model but also achieves better performance than transformer-based models [35]. Evaluation scores of different metrics on the BanCap dataset for different models along with the proposed methodology under a number of categories are shown in following "Figure 5", for easy comparison between them. Single encoder and single decoder are

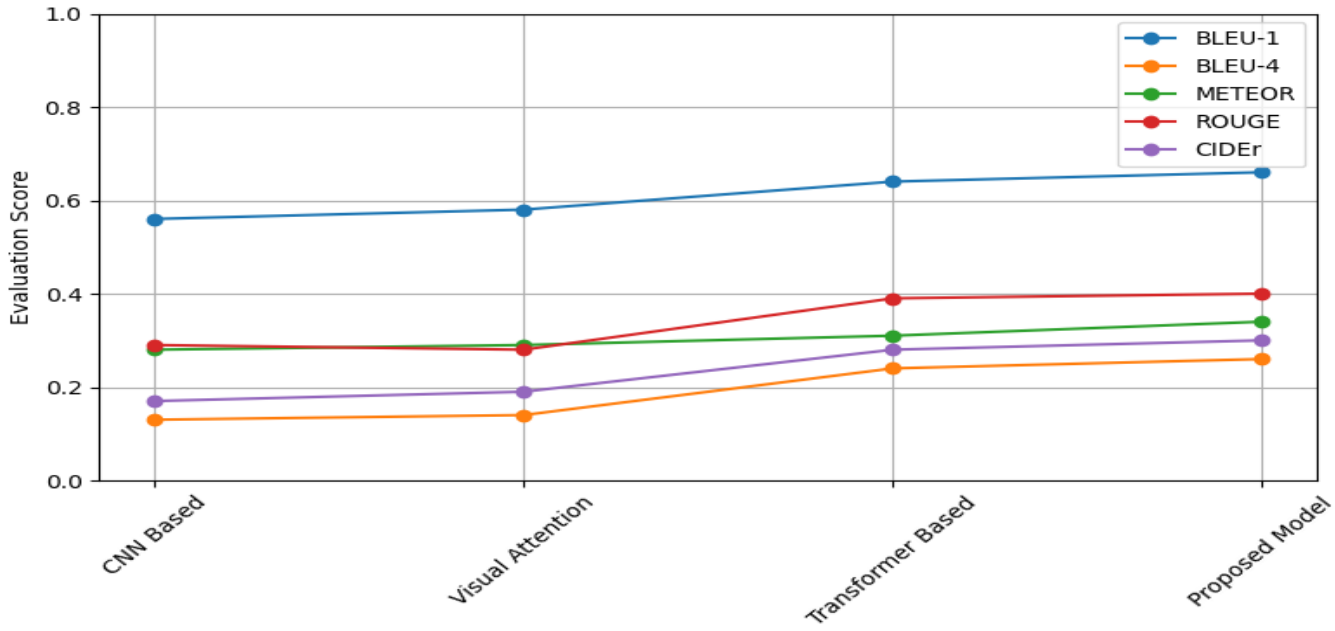

Figure 5. Summary of evaluation score obtained from several models.

used in the transformer-based decoder of the proposed hybrid model, reducing the number of learnable parameters in comparison to [29, 35]. The transformer-based image captioning model discussed in [29] took more than twelve hours and the model in [35] requires more than nine hours to train the model, whereas exploring our proposed model using distinct CNN requires around four hours for training. Since for image analysis, pre-trained models are used, the hybrid captioning model training complexity and efficiency mainly depend upon the decoding component. "Table 3", shows the number of parameters and floating-point

Table 3. Parameters and operations for linguistic decoding.

Model	Sequence Length	Attention Head	No of Encoder /decoder	Number of FLOPs (billion)	Number of Parameters (million)
Transformer Based Image captioning in Bengali [35]	25	8	8	2.139	25.174
Transformer Based Image captioning in Hindi [29]	34	8	6	2.196	18.881
Proposed Model	34	2	1	0.145	3.147

operations required every second to perform linguistic decoding for the proposed model and other Hindi and Bengali Image captioning systems based on transformer-based models. The model depicted in [29] requires more than 18 million parameters and 2.196 billion FLOPs, while [35] requires 25.174 million parameters and 2.139 billion FLOPs to perform decoding. Despite reducing the number of required parameters to 3.147 million and the need for a number of operations to 0.145 billion FLOPs in the decoder, the proposed model ensures accurateness in training with the BanCap dataset due to the incorporation of advanced CNN mechanism, along with improved activation patterns, regularization mechanism, and advanced optimization techniques. The proposed model using EfficientNetV2S requires 21.6 million parameters and 8.8 billion FLOPs for image analysis while ConvNextSmall and InceptionResnetV2 require 50.2 and 55.9 million parameters and 8.7 and 6.65 billion FLOPs respectively. Thus the proposed model using EfficientNetV2S provides the best trade-off between computational requirement, parameter count, and evaluation score of different performance metrics, making it the most efficient choice for the proposed Bengali Image Captioning framework. The loss graph depicted in "Figure 3" for the proposed model using three different types of CNN shows the training and validation loss remain high initially. However, with the progression of iterations, both the losses gradually diminish and steadily approach towards convergence. Hence, the proposed model demonstrates effective learning, ensuring robustness across visual inputs, and avoids overfitting to the training data. Despite differences in statistical properties of feature representation techniques in the proposed encoder, the learning dynamics and loss values of the proposed model using three CNNs exhibit similar trends towards convergence as the proposed hybrid model trained under identical conditions, including the same dataset and data preprocessing pipeline, optimization technique, loss function, and other hyperparameters along with the same improvised transformer in the proposed decoder. However, the most remarkable observation from "Figure 4(a) and 4(b)", is the obtained descriptions are found to be very lively and even more scene-specific and informative than the real descriptions.

6.1. Analysis of Visual Examples

The proposed model incorporates both spatial and semantic information effectively by attending to different parts of the image and leveraging pre-trained visual representations, the predicted descriptions are more contextually rich and semantically accurate in comparison to other attention-based approaches. Visualization of the attention plot is done in this context.

Figure 6. Visualization of Attention Plot.

"Figure 6" entails that focus on the image region is properly aligned with the generated word. Despite using complex multimodal models [10], full attention-based models [11], or any explicit object detection [28] techniques, the number of primary objects, as well as the relation between them concerning the background is well found in the descriptions. While in the first two images, the number of person and boy is "two" which is called "দুজন" or "দুই" in Bengali, the second two pictures contain one person, which in Bengali is "এক". Not only are the descriptions appropriate to the visuals, but they also contain some variation while the main content remains the same. For example, one of the descriptions of the first picture says, "Two men stand outside wearing jackets and hats", whereas the other describes "Two men are standing on a city street at night". Thus, more information is found by exploring the images with distinct CNN in the proposed framework. However, the ground truth expresses that "Two persons are posing in front of the camera in the night", but since no "camera" is found in the visuals and our model is learning instead of memorizing, it predicts without any information of the camera. The "water", "snow", and "mountain" are properly figured out in the successive image descriptions. However, in the fourth image, even the "dog with its colour" is detected, but the "car" in the background is missing due to a lack of supervised information in training data. In all the predicted descriptions of the fifth image, "সমুদ্রের তীরে" or "সমুদ্রের ধারে" or " সৈকতে" and "বালিতে লিখছে" is found which signifies the happening spot and purpose remains the same, however from the different descriptions the additional information found is "খালি গায়ে", " লাঠি দিয়ে" and "পাথুরে সমুদ্রের ধারে" which means "they were naked" and "using stick" to write in sand aside "stone filled sea beach". From the sixth image, the obtained distinct information is "the dogs are playing on the sea beach" and "three dogs are chasing a ball thrown in the sea". Although the ground truth tells the number of dogs and their destination, the model-generated descriptions indicate the purpose of chasing the ball more clearly. The main content of the next visual is "a girl is sliding", while the obtained additional information from different descriptions is the girl is "little" and wearing a "white dress" and sliding in a "blue" colour slide in "the playground". The ground truth of the last visual is "সাদা জামা ও চামড়ার কোট পরা এক লোক" which means "A person wearing a white shirt and leather jacket" while the common theme in predicted descriptions is "a person looking downwards" and additional information found in the descriptions are "কৃষ্ণ বর্ণের লোক", "রোদচশমা" which means "the person is black" and wearing "sunglass" along with a white shirt and leather jacket. Hence, the description produced by the proposed models from different perceptions is not only appropriate to the visuals but also more informative and enriched compared to the ground truth. Despite finding more objects for some images, such as the first and fourth images on applying heatmap based on object detection and overlay as visualized in below "Figure 7", the proposed framework identifies primary objects and

Figure 7. Heatmap based on object detection and overlay on the original image.

significant components in the visuals, like person, gender, dog, sea, mountain, slide, nighttime etc., providing a comprehensive overview of the scene. The descriptions follow a logical pattern of human understanding, describing the most significant component first and then moving to the rest. It is found that semantically appropriate descriptions are obtained by establishing the relationships between those objects. Thus inside activity is properly recognized, and produced descriptions are more specific to the visual content. Additionally, the proposed model recognizes the number and colour of significant objects properly and produces a vibrant description. While distinct image analysis

techniques in the proposed framework provide multiple vivid descriptions, providing additional information for the same visual, it lacks specific details such as the mental state of the people or the season or location of the spotted visuals, the exact colour variations in the sky, or seawater etc. The model tends to provide factual descriptions without probing into emotional responses or interpretive insights about the scene, such as tranquillity or a sense of vastness and does not capture the dynamic aspects of the scene, such as the movement of people or the possible noise and activity levels. Despite producing contextually appropriate descriptions, the model sometimes misses broader contextual information that could add depth to the description, such as cultural, historical, or environmental context focusing on that region. For ease of comparison with the GPT-based model, the descriptions produced in English language using "Image Caption Generator" are translated into Bengali and presented along with the proposed model descriptions for every visual in "Figure 4 (a) and 4 (b)". Analysing these descriptions, it is found that while the GPT-based model tends to produce more fluent and natural descriptions, the proposed framework tends to produce specific and additional information utilizing distinct image analysis mechanisms. In the GPT-based description of the first visual, the mental state of the person i,e; "good vives" or "ভাল মেজাজে" is listed, whereas the proposed framework describes their clothes, "জ্যাকেট এবং টুপি". The GPT description of the second and seventh pictures entails the "splashing" and "sliding" time is summer or "গ্রীষ্মে" as well as "মজার", "আনন্দে" indicates their happy mental state. However, it misses two boys' splashing spots and the little girl's sliding details, dress colour etc. Similarly, in the GPT-based third image description, "অ্যাড্রেনালিন রাশ" clearly tells the person in the parachute is stressed, whereas "আরামদায়ক দিন উপভোগ" in the fourth description indicates the woman is enjoying the company of her pet friend, but it lacks description of the pet, which was found using the proposed model. The fifth caption hints that the boys on the sea beach are competing for memory, which is termed "স্মৃতি" in Bengali. Despite providing natural descriptions in a broader context, the GPT models did not describe the sea beach or description of the boys. In the sixth description, the GPT model says that the dogs were chasing sea waves, whereas the proposed framework tells more specifically that the number of dogs was three, and they were chasing a ball thrown on the sea. GPT-generated description of the last visual indicates that the person is enjoying the city vibes in style. Although it is fluent and attractive, it lacks information regarding the person, like his colour and clothing, which was well obtained using the proposed framework.

Conclusion

This paper presents a novel approach to generating Bengali descriptions by utilizing a diverse range of CNNs, including ConvNext, EfficientNet, and Inception ResNet, combined with a Transformer-based decoder model. The study focuses on the feasibility of creating Bengali captions for visually impaired individuals and multitasking users in Bengali-speaking regions. The model leverages various CNN variants to enhance feature representation without increasing computational complexity. It produces context-specific and diverse Bengali descriptions, both in text and audio formats, that are richer and more informative than human-annotated ground truths. The proposed framework obtains BLEU scores of 0.66 (unigram), 0.59 (bigram), 0.44 (trigram), and 0.26 (four-gram) and METEOR, ROUGE, CIDEr scores of 0.34, 0.40, 0.30 respectively utilizing EfficientNetV2S and improvised Transformer in the encoder-decoder framework, outperforms state-of-the-art models in evaluations. However, some limitations are noted, including the risk of bias if training images are presented in a fixed order, which could affect caption generation. Additionally, the paper assumes compatibility between the Bengali captions from the BanCap dataset and the images from the Flickr 8k dataset, despite unknown compatibility, leading to slightly higher training loss compared to validation loss.

While the proposed model performs well with common objects, it may struggle with more diverse, complex, and varied content. To create a more robust and generalized model, it is essential to train with larger datasets that include complex scenes and enriched Bengali descriptions reflecting various cultural contexts, urban and rural settings, and different socio-economic backgrounds. Data augmentation can help mitigate the limitations of local and regional datasets. With access to larger and more diverse datasets, a more versatile system could be developed using a GPT based system or an end-to-end transformer-based approach. Ensembling different image analysis techniques or integrating object detection mechanisms could enhance captioning performance, providing more accurate, compact descriptions that include every visual detail, especially for complex images with multiple objects. Captioning system based on specific domain dataset enhances user experience in personalized description generation. To replicate the study for other low-resource languages, modification of the text tokenization process can be done according to the syntax and grammar of the target language along with adjustment of model parameters, better captures linguistic nuances and generates proper descriptions in the targeted language.

Author contributions

Biswajit Patra: Conceptualization, Methodology, Data preparation, Software, Validation, Visualization, Designing and Drafting of the original manuscript. **Dakshina Ranjan Kisku:** Conceptualization, Methodology, Investigation, Supervision and Draft correction.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).
2. Gülgün, O. D., & Erol, H. (2020). Classification performance comparisons of deep learning models in pneumonia diagnosis using chest x-ray images. Turkish Journal of Engineering, 4(3), 129-141. https://doi.org/10.31127/tuje.652358
3. Atalay Aydın, V. (2024). Comparison of CNN-based methods for yoga pose classification. Turkish Journal of Engineering, 8(1), 65-75. https://doi.org/10.31127/tuje.1275826
4. Pajaziti, A., Basholli, F., & Zhaveli, Y. (2023). Identification and classification of fruits through robotic system by using artificial intelligence. Engineering Applications, 2(2), 154-163.
5. Meghraoui, K., Sebari, I., Bensiali, S., & El Kadi, K. A. (2022). On behalf of an intelligent approach based on 3D CNN and multimodal remote sensing data for precise crop yield estimation: Case study of wheat in Morocco. Advanced Engineering Science, 2, 118-126.
6. Sıngh, S., Kumar, K., & Kumar, B. (2024). Analysis of feature extraction techniques for sentiment analysis of tweets. Turkish Journal of Engineering, 8(4), 741-753. https://doi.org/10.31127/tuje.1477502
7. Othman, M. M. (2023). Modeling of daily groundwater level using deep learning neural networks. Turkish Journal of Engineering, 7(4), 331-337. https://doi.org/10.31127/tuje.1169908
8. Sen, O., Fuad, M., Islam, M. N., Rabbi, J., Masud, M., Hasan, M. K., ... & Iftee, M. A. R. (2022). Bangla natural language processing: A comprehensive analysis of classical, machine learning, and deep learning-based methods. IEEE Access, 10, 38999-39044.
9. Patra, B., & Kisku, D. R. (2023, December). Precise and Faster Image Description Generation with Limited Resources Using an Improved Hybrid Deep Model. In International Conference on Pattern Recognition and Machine Intelligence (pp. 166-175). Cham: Springer Nature Switzerland.
10. Yu, J., Li, J., Yu, Z., & Huang, Q. (2019). Multimodal transformer with multi-view visual representation for image captioning. IEEE transactions on circuits and systems for video technology, 30(12), 4467-4480.
11. Parvin, H., Naghsh-Nilchi, A. R., & Mohammadi, H. M. (2023). Image captioning using transformer-based double attention network. Engineering Applications of Artificial Intelligence, 125, 106545.
12. [Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
13. Miller, A., Fisch, A., Dodge, J., Karimi, A. H., Bordes, A., & Weston, J. (2016). Key-value memory networks for directly reading documents. arXiv preprint arXiv:1606.03126.
14. Li, Z., Li, Y., & Lu, H. (2019). Improve image captioning by self-attention. In Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part V 26 (pp. 91-98). Springer International Publishing.
15. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057). PMLR.
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
17. Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10578-10587).
18. Vasireddy, I., HimaBindu, G., & Ratnamala, B. (2023). Transformative Fusion: Vision Transformers and GPT-2 Unleashing New Frontiers in Image Captioning within Image Processing. International Journal of Innovative Research in Engineering & Management, 10(6), 55-59.
19. Mishra, S., Seth, S., Jain, S., Pant, V., Parikh, J., Jain, R., & Islam, S. M. (2024, May). Image Caption Generation using Vision Transformer and GPT Architecture. In 2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT) (pp. 1-6). IEEE.
20. Hu, Y., Hua, H., Yang, Z., Shi, W., Smith, N. A., & Luo, J. (2023). Promptcap: Prompt-guided image captioning for vqa with gpt-3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2963-2975).
21. Kurlekar, S., Deshpande, O., Kamble, A., Omanna, A., & Patil, D. (2020). Reading Device for Blind People using Python OCR and GTTS. International Journal of Science and Engineering Applications, 9(4), 049-052.
22. Granquist, C., Sun, S. Y., Montezuma, S. R., Tran, T. M., Gage, R., & Legge, G. E. (2021). Evaluation and Comparison of Artificial Intelligence Vision Aids: Orcam MyEye 1 and Seeing AI. Journal of Visual Impairment & Blindness, 115(4), 277-285. https://doi.org/10.1177/0145482X211027492
23. Coughlan, J. M., & Miele, J. (2017, October). AR4VI: AR as an accessibility tool for people with visual impairments. In 2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct) (pp. 288-292). IEEE.
24. Wu, S., Wieland, J., Farivar, O., & Schiller, J. (2017, February). Automatic alt-text: Computer-generated

image descriptions for blind users on a social network service. In proceedings of the 2017 ACM conference on computer supported cooperative work and social computing (pp. 1180-1192).
25. Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., ... & Bigham, J. P. (2018). Vizwiz grand challenge: Answering visual questions from blind people. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3608-3617).
26. Doore, S. A., Istrati, D., Xu, C., Qiu, Y., Sarrazin, A., & Giudice, N. A. (2024). Images, Words, and Imagination: Accessible Descriptions to Support Blind and Low Vision Art Exploration and Engagement. Journal of Imaging, 10(1), 26.
27. Shrestha, R. (2022, February). A transformer-based deep learning model for evaluation of accessibility of image descriptions. In Proceedings of the 2022 14th International Conference on Machine Learning and Computing (pp. 28-33).
28. Mishra, S. K., Harshit, Saha, S., & Bhattacharyya, P. (2022). An Object Localization-based Dense Image Captioning Framework in Hindi. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(2), 1-15.
29. Mishra, S. K., Dhir, R., Saha, S., Bhattacharyya, P., & Singh, A. K. (2021). Image captioning in Hindi language using transformer networks. Computers & Electrical Engineering, 92, 107114.
30. Rahman, M., Mohammed, N., Mansoor, N., & Momen, S. (2019). Chittron: An automatic bangla image captioning system. Procedia Computer Science, 154, 636-642.
31. Deb, T., Ali, M. Z. A., Bhowmik, S., Firoze, A., Ahmed, S. S., Tahmeed, M. A., ... & Rahman, R. M. (2019). Oboyob: A sequential-semantic bengali image captioning engine. Journal of Intelligent & Fuzzy Systems, 37(6), 7427-7439.
32. Al Faraby, H., Azad, M. M., Fedous, M. R., & Morol, M. K. (2020, December). Image to Bengali caption generation using deep CNN and bidirectional gated recurrent unit. In 2020 23rd international conference on computer and information technology (ICCIT) (pp. 1-6). IEEE.
33. Humaira, M., Shimul, P., Jim, M. A. R. K., Ami, A. S., & Shah, F. M. (2021). A hybridized deep learning method for bengali image captioning. International Journal of Advanced Computer Science and Applications, 12(2).
34. Ami, A. S., Humaira, M., Jim, M. A. R. K., Paul, S., & Shah, F. M. (2020, December). Bengali image captioning with visual attention. In 2020 23rd International Conference on Computer and Information Technology (ICCIT) (pp. 1-5). IEEE.
35. Hossain, M. A., Hasan, M. A. R., Hossen, E., Asraful, M., Faruk, M. O., Abadin, A. Z., & Ali, M. S. Automatic Bangla Image Captioning Based on Transformer Model in Deep Learning.
36. Tan, M., & Le, Q. (2021, July). Efficientnetv2: Smaller models and faster training. In International conference on machine learning (pp. 10096-10106). PMLR.
37. Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11976-11986).
38. Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1).
39. Khan, M. F., Shifath, S. M., & Islam, M. S. (2022). BAN-cap: a multi-purpose English-Bangla image descriptions dataset. arXiv preprint arXiv:2205.14462.
40. Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415.
41. Chen, X., Kang, B., Wang, D., Li, D., & Lu, H. (2022, October). Efficient visual tracking via hierarchical cross-attention transformer. In European Conference on Computer Vision (pp. 461-477). Cham: Springer Nature Switzerland.
42. Arafat, M. Y., Fahrin, S., Islam, M. J., Siddiquee, M. A., Khan, A., Kotwal, M. R. A., & Huda, M. N. (2014, December). Speech synthesis for bangla text to speech conversion. In The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014) (pp. 1-6). IEEE.
43. Terven, J., Cordova-Esparza, D. M., Ramirez-Pedraza, A., & Chavez-Urbiola, E. A. (2023). Loss functions and metrics in deep learning. A review. arXiv preprint arXiv:2307.02694.
44. Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.
45. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).
46. Denkowski, M., & Lavie, A. (2014, June). Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the ninth workshop on statistical machine translation (pp. 376-380).
47. Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).
48. Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4566-4575)