

VisionXAI: An Image Captioning System with Explainable AI

January 2023

Prepared by:

Abu Musa Sakib
ID: 2321673

Sumyia Afnan Mukta
ID: 2321672

Department of Computer Science and Engineering
Independent University, Bangladesh

Supervised by

Ashraful Islam, PhD
Associate Professor
Department of Computer Science and Engineering
Independent University, Bangladesh

Attestation

We are aware of the fact that plagiarism is strictly prohibited and it conflicts with our university's rules and regulations. We guarantee the authenticity of our work. We have used some copyrighted material and models of others in our project which has been properly cited following the international standards proper guidelines.

Author Name:

. . . . . .

Signature:

. . . . . .

Author Name:

. . . . . .

Signature:

. . . . . .

Author Name:

. . . . . .

Signature:

. . . . . .

Evaluation Committee

Supervisor

Name: _____ Signature: _____

Internal Examiner 1

Name: _____ Signature: _____

Internal Examiner 2

Name: _____ Signature: _____

External Examiner

Name: _____ Signature: _____

Acknowledgement

We would like to express sincere gratitude to our respected Professor, AKM Mahbubur Rahman, Independent University, Bangladesh. His dedicated guidance, insightful suggestions, remarkable patience, and the generous investment of his valuable time played a pivotal role in the successful completion of our project. Our Professor's expertise and mentorship greatly enhanced the quality and depth of our work, and we are truly appreciative of his unwavering support throughout our academic journey.

Additionally, our heartfelt thanks go to the CCDS Lab (Center for Computational & Data Sciences) for their extraordinary help and consistent support. The collaborative and resourceful environment provided by CCDS Lab significantly contributed to the development and execution of our project. Moreover, The encouragement and assistance have been instrumental in shaping our academic and research field. We are grateful for the opportunities and resources extended to us during our research, which undoubtedly enriched our learning experience.

We acknowledge and appreciate the contributions of all individuals who supported us, both directly and indirectly, during this endeavor.

Abstract

Automated image captioning seeks to generate human-like descriptions of images by combining computer vision and natural language processing (NLP). Convolutional neural networks (CNNs) encode visual features, while recurrent neural networks (RNNs) or transformers decode them into text, bridging the gap between visual and linguistic understanding. While effective, such models often lack transparency. Explainable AI (xAI) addresses this challenge by highlighting how image regions influence generated captions, improving interpretability and user trust. In this work, we focus on Bengali image captioning and present VisionXAI, a mobile app that generates captions with text-to-speech (TTS) support for visually impaired users. By integrating xAI, the system not only produces captions but also explains its reasoning, ensuring accessibility, reliability, and inclusiveness in real-world use.

Contents

1 Introduction **8**
- 1.1 Background of the project . . . 8
- 1.2 Problem statement . . . 8
- 1.3 Outcome of the project . . . 8
  - 1.3.1 Contributions . . . 9

2 Literature Review **10**
- 2.1 Systematic Literature Review Methodology . . . 10
  - 2.1.1 Search Strategy . . . 10
  - 2.1.2 Inclusion and Exclusion Criteria . . . 10
  - 2.1.3 Study Selection and Analysis . . . 10
- 2.2 Foundations of image captioning . . . 10
- 2.3 Explainable image captioning (xAI) . . . 12
- 2.4 Bengali image captioning and low-resource challenges . . . 12
- 2.5 Relevant NLP and XAI work informing VisionXAI . . . 12
- 2.6 Gaps and opportunities . . . 13

3 Introducing the Image Captioning Dataset **14**
- 3.1 Motivation & Purpose . . . 14
- 3.2 Source of Data Samples . . . 14
- 3.3 Structure and Features of the Dataset . . . 15
  - 3.3.1 Diversity of Visual Categories . . . 15
  - 3.3.2 Caption Characteristics . . . 15
  - 3.3.3 Linguistic Richness and Relevance . . . 16
  - 3.3.4 Annotation Reliability . . . 16
- 3.4 Dataset Summary . . . 16
- 3.5 Model . . . 16
  - 3.5.1 Model Training . . . 16
  - 3.5.2 Model Architecture . . . 17
  - 3.5.3 Training Algorithm . . . 19
- 3.6 Discussion . . . 19
  - 3.6.1 Epoch vs. Loss Behavior . . . 19
  - 3.6.2 Model Evaluation Results . . . 20

4 Results and Analysis **22**
- 4.1 Overview . . . 22
- 4.2 EfficientNet-B4 encoder + Transformer decoder . . . 22
- 4.3 InceptionV3 encoder + Bahdanau attention + GRU decoder . . . 23
- 4.4 Qualitative outputs and attention . . . 23
- 4.5 Comparative analysis and caveats . . . 23

4.5.1 Convergence . . . 24
4.5.2 Stability . . . 24
4.5.3 Interpretability . . . 24
4.5.4 Decoding behavior and its effects . . . 24
4.5.5 Limitations of Comparison . . . 24
4.6 Summary . . . 25

5 Challenges, Future Works, and Conclusion **27**
5.1 Challenges . . . 27
5.1.1 Dataset Size and Memory Constraints . . . 27
5.1.2 Model Complexity and Training Time . . . 27
5.1.3 Hyperparameter Tuning and Optimization . . . 27
5.1.4 Caption Quality and Model Performance . . . 28
5.1.5 Dataset Quality and Quantity Issues . . . 28
5.1.6 Training Environment Limitations . . . 28
5.1.7 Dataset and Result Bias . . . 28
5.1.8 Language-Specific Challenges . . . 29
5.2 Future Work . . . 29
5.2.1 Advancing Model Architectures . . . 29
5.2.2 Improving Attention, Alignment, and Localization . . . 29
5.2.3 Dataset Expansion and Annotation Improvements . . . 29
5.2.4 Leveraging Pretraining and Transfer Learning . . . 29
5.2.5 Improving Training Efficiency . . . 30
5.2.6 Mitigating Dataset and Model Bias . . . 30
5.2.7 Enhancing Caption Fluency . . . 30
5.2.8 Developing APIs for Organizational Integration . . . 30
5.3 Conclusion . . . 30

Bibliography **31**

List of Figures

1 Encoder–Decoder Architecture with Bahdanau Attention for Bengali Image Captioning 18
2 Training and validation loss across epochs for the image captioning model. 20

1 Introduction

1.1 Background of the project

Automated image captioning combines advances in computer vision and natural language processing to generate human-like descriptions of visual content. Modern captioning systems typically use convolutional neural networks (CNNs) to encode images and sequence models (e.g., recurrent networks or transformer decoders) to produce fluent natural language descriptions. By bridging visual understanding and language generation, these systems enable a wide range of applications from image search and organization to accessibility aids for visually impaired users.

Alongside accuracy, reliability and transparency have become critical requirements for real-world deployment. Explainable AI (xAI) techniques provide insight into which image regions or model components drive particular words in a caption, improving user trust and facilitating debugging. For assistive applications, combining captioning with text-to-speech (TTS) makes visual content accessible to people with vision loss; exposing the model's reasoning further increases user confidence in the generated descriptions.

1.2 Problem statement

Despite strong progress in captioning benchmarks, producing trustworthy, localized captions in languages other than English remains challenging. Off-the-shelf captioners are often trained on large English datasets and do not transfer cleanly to other languages or to resource-constrained mobile deployments. In addition, purely black-box caption outputs make it hard for users and developers to validate or correct mistakes. For an accessible mobile app, we therefore need a system that produces accurate Bengali captions, speaks them aloud with TTS, and gives interpretable explanations of why a caption was produced.

1.3 Outcome of the project

This project implements VisionXAI, an image captioning system that focuses on three goals:

- Produce high-quality Bengali image captions by adapting modern encoder–decoder architectures to Bengali text generation and localized linguistic patterns.
- Provide concise, human-readable explanations that highlight image regions and model evidence supporting individual words or phrases (xAI components).
- Integrate caption generation with on-device text-to-speech so that captions can be read aloud for accessibility on mobile devices.

These goals emphasize practical deployment: compact models or pruning/quantization strategies for mobile inference, robust preprocessing for images captured in everyday settings, and careful handling of language-specific tokenization and morphology when generating Bengali text.

1.3.1 Contributions

The main contributions of this work are:

- A pipeline for Bengali image captioning that adapts modern vision–language architectures to generate natural Bengali captions and integrates a TTS component for spoken output.
- An xAI module that links generated caption tokens to image regions and visualizes these associations, improving transparency for end users and developers.
- Implementation and evaluation on representative images, showing qualitative and quantitative improvements over baseline captioning approaches when localized to Bengali and when explanation feedback is available.
- Practical guidance for on-device deployment, including performance/latency measurements and recommended model compression strategies for mobile usage.

2 Literature Review

This chapter reviews prior work on automatic image captioning, explainable image captioning (xAI), and Bengali-language captioning and language resources. It summarizes the main architectural trends, common datasets, and evaluation practices, then highlights gaps that motivate VisionXAI.

2.1 Systematic Literature Review Methodology

A systematic literature review (SLR) was conducted following PRISMA-style guidelines to ensure comprehensive, transparent, and reproducible coverage of prior research relevant to Bengali image captioning and explainable vision–language systems. The review focused on peer-reviewed journal articles, conference proceedings, and high-impact arXiv preprints published between 2018 and 2024.

2.1.1 Search Strategy

Literature was retrieved using academic databases including IEEE Xplore, ACM Digital Library, SpringerLink, and arXiv. Search queries combined keywords such as *image captioning*, *Bengali image captioning*, *vision–language models*, *explainable AI*, *attention mechanisms*, and *contrastive learning*.

2.1.2 Inclusion and Exclusion Criteria

Studies were included if they (i) proposed or evaluated image captioning models, (ii) addressed Bengali or other low-resource languages, or (iii) incorporated explainability or vision–language alignment techniques. Works focusing exclusively on English captioning without transferable methodological contributions were excluded.

2.1.3 Study Selection and Analysis

After removing duplicates, titles and abstracts were screened for relevance, followed by full-text review. Selected studies were analyzed with respect to dataset characteristics, model architectures, explainability mechanisms, evaluation metrics, and practical deployment considerations. The final set of studies is summarized in Table 2.1, highlighting how existing research informs and motivates the design choices of VisionXAI.

2.2 Foundations of image captioning

Automatic image captioning sits at the intersection of computer vision and natural language generation. Early encoder–decoder approaches used a CNN encoder to extract

Table 2.1: Summary of Key Studies in Bengali Image Captioning and Explainable Vision–Language Modeling

Study	**Core Method**	**Main Contribution**	**Relevance to VisionXAI**
Chittron [22]	VGG16 + LSTM	First culturally grounded Bangla captioning system; qualitative relevance despite low BLEU	Motivates Bengali-specific datasets and cultural grounding
TextMage [15]	CNN + LSTM + Attention	Reduced Western bias using BanglaLekhaImageCaptions	Supports dataset localization and attention-based decoding
CNN–BiGRU [1]	CNN + BiGRU	Efficient sequential modeling for small Bengali datasets	Informs lightweight decoder selection
CNN–Transformer [21]	CNN + Transformer	Improved long-range dependency modeling	Justifies Transformer decoder in VisionXAI
EfficientNet–Transformer [14]	EfficientNet + Transformer	State-of-the-art Bengali captioning scores	Supports hybrid encoder–decoder design
Context-Aware Attention [5]	ResNet + BiGRU + Attention	Strong gains via sub-region focus	Guides explainable attention visualization
Explainable CNN–CNN [19]	CNN Encoder–Decoder + Hierarchical Attention	Region-word alignment without RNNs	Informs interpretable architecture choices
VALE [20]	XAI + Vision–Language Models	Combined visual and textual explanations	Aligns with multimodal explanation goals
CoCa [30]	ViT + Contrastive + Generative Learning	Unified foundation vision–language model	Inspires contrastive pretraining direction

visual features and an RNN decoder (LSTM/GRU) to generate captions [28, 11, 16]. Attention mechanisms that align words to image regions improved both quality and interpretability, notably in the "Show, Attend and Tell" framework [29]. More recently, transformer architectures and contrastive pretraining have further advanced visual–language modelling by improving long-range dependencies and representation alignment [27, 7].

Representative developments include:

- CNN–RNN models with attention that localize image regions for each generated token [29].
- Pure transformer or hybrid encoder–decoder models that leverage self-attention for both vision and language processing [27].
- Contrastive pretraining and multimodal foundation models that scale well and transfer across tasks [6].

2.3 Explainable image captioning (xAI)

Explainability in captioning aims to show which image regions or model components justify individual words or phrases. Visual attribution methods such as Grad-CAM provide localization maps that can be paired with generated caption tokens to produce human-readable explanations [24]. Model-agnostic explanation techniques like LIME and SHAP give complementary, example-based explanations and feature importance summaries [23, 18].

Recent xAI captioning systems combine visual attribution with textual rationales to produce multimodal explanations (visual highlights + short Bengali text explaining why a word was produced). These approaches improve user trust and facilitate error analysis, particularly in assistive contexts where users rely on the caption for situational awareness.

2.4 Bengali image captioning and low-resource challenges

Most large captioning datasets and models are developed for English (e.g., MS COCO). For Bengali, the literature is smaller but growing: several works adapt encoder–decoder models (CNN+RNN or transformer-based) and report qualitative gains when training on Bengali caption datasets (e.g., BanglaLekhaImageCaptions and related corpora) [4, 3]. Practical constraints include limited paired data, spelling variation, and morphology that affects tokenization.

Proposed strategies for Bengali captioning include strong transfer learning from large multilingual or vision–language models, careful tokenization and subword modeling for Bengali morphology, and data augmentation to increase caption diversity. The literature shows that attention mechanisms and multimodal pretraining are especially helpful for modeling languages with limited supervised data [27, 7].

2.5 Relevant NLP and XAI work informing VisionXAI

Although much prior work focuses on English, techniques translate to Bengali with careful localization. Transformer decoders, beam search, and hierarchical attention mechanisms

are commonly used for higher-quality captions. For explainability, combining Grad-CAM style visual maps with token-level attribution provides compact, actionable explanations for end users [29, 24].

In the Bangla NLP space, efforts such as BanglaBERT and BanglaT5 demonstrate the benefits of monolingual pretraining and targeted generation models for tasks in Bengali [2, 3]. Those resources inform our choice of language models and tokenizer design when generating Bengali captions.

2.6 Gaps and opportunities

Key gaps motivating VisionXAI are:

- Limited publicly available, high-quality Bengali image caption corpora with diverse, colloquial captions.
- Few systems that jointly optimize caption quality, explicit visual explanations, and on-device speech output for accessibility.
- Practical engineering guidance for deploying compact captioning + xAI + TTS on mobile devices (latency, pruning/quantization strategies).

VisionXAI addresses these gaps by adapting modern vision–language and xAI techniques to Bengali, integrating an on-device TTS pipeline, and evaluating both caption quality and explanation usefulness on representative images.

Table 2.1 synthesizes representative studies across Bengali image captioning, explainable AI, and modern vision–language modeling. While prior work demonstrates steady progress in caption quality and interpretability, most systems focus on isolated aspects such as caption accuracy, attention visualization, or multilingual transfer. In contrast, VisionXAI integrates these strands by combining a Bengali-adapted vision–language architecture with explicit visual explanations and an on-device text-to-speech pipeline. This holistic design directly addresses accessibility, transparency, and deployment constraints that remain underexplored in existing literature.

3 Introducing the Image Captioning Dataset

3.1 Motivation & Purpose

The development of a comprehensive dataset for Bangla image captioning emerges from a critical need to advance multimodal learning capabilities in the Bangla language. Existing global datasets tend to be dominated by English captions or culturally distant imagery, resulting in a noticeable gap when attempting to build models capable of understanding and describing visuals in a Bangla context. While several Bangla-specific datasets exist, many are either limited in scope, domain, or vocabulary, and often fail to encapsulate the visual diversity encountered in real-world scenarios across Bangladesh.

Recognizing these limitations, our work focuses on assembling a versatile and enriched dataset that accommodates a wide range of image categories—including humans, animals, objects, natural scenes, and daily activities. This initiative is driven by the intent to construct a resource that not only supports effective image captioning but also reflects the cultural and environmental nuances unique to Bangla-speaking regions.

Beyond serving as a training corpus, this dataset represents a deeper commitment to fostering advancements in Bangla-language AI research. By providing a diverse and linguistically rich set of image–caption pairs, we aim to empower future work in computer vision, natural language processing, and multimodal systems tailored specifically for Bangla users.

3.2 Source of Data Samples

The dataset comprises a total of 18,542 images, derived from multiple established collections such as BAN-Cap, BNATURE, and BanglaLekhaImageCaptions, alongside an additional custom dataset created to enrich underrepresented categories. Each of these sources offers distinctive characteristics:

- **BAN-Cap** contributes human-centric and activity-based images paired with Bangla captions.
- **BNATURE** offers rich visuals of wildlife, landscapes, and natural environments.
- **BanglaLekhaImageCaptions** provides objects, daily-use items, and contextual indoor scenes.
- **The custom dataset** fills the gaps in categories not adequately represented in existing datasets, ensuring broad coverage and maximizing diversity.

The selection of these sources is both intentional and strategic. Together, they form a multifaceted representation of Bangladeshi visual contexts, ensuring that the resulting dataset avoids domain imbalance and mirrors the visual variety inherent in day-to-day life.

Each image is accompanied by one or more human-annotated descriptions, ensuring that the dataset maintains linguistic quality, clarity, and contextual accuracy. These captioning efforts were guided by established standards in Bangla grammar and style, ensuring caption uniformity and reliability.

3.3 Structure and Features of the Dataset

The image captioning dataset is meticulously structured to support a wide range of machine learning tasks, particularly those involving vision–language understanding. The key features are summarized below:

3.3.1 Diversity of Visual Categories

The dataset spans numerous categories, including but not limited to:

- Humans and group interactions
- Domestic and wild animals
- Everyday objects
- Urban and rural environments
- Nature and scenic landscapes

This broad representation ensures that models trained on the dataset develop the ability to generalize across varied real-world scenes.

3.3.2 Caption Characteristics

The captions associated with each image provide detailed Bangla descriptions crafted to reflect:

- Clear object identification
- Contextual understanding of activities
- Spatial relationships
- Scene-level interpretations

With a vocabulary size of 22,276 words and a maximum caption length of 38 words, the dataset offers an expansive linguistic range, supporting both simple and complex caption generation tasks.

3.3.3 Linguistic Richness and Relevance

The dataset incorporates:

- Formal and semi-formal descriptive expressions
- Natural sentence structures
- Bangla-specific linguistic nuances

This ensures that the resulting captioning models are capable of producing fluent, meaningful, and contextually appropriate Bangla descriptions.

3.3.4 Annotation Reliability

To maintain consistency and linguistic accuracy:

- Captions were validated using recognized Bangla language resources.
- Multiple annotators ensured clarity and uniformity in descriptions.
- Quality control measures were applied to avoid ambiguity or redundancy.

3.4 Dataset Summary

Collectively, the dataset serves as a high-quality resource for training, evaluating, and benchmarking Bangla image captioning models. Its balanced composition of diverse images paired with descriptive Bangla captions provides a robust foundation for research in multimodal AI. By capturing the intricacies of both visual contexts and linguistic expression, this dataset contributes significantly to advancing Bangla-language computer vision technologies and supports future explorations in image understanding, caption generation, and cross-modal learning.

3.5 Model

3.5.1 Model Training

The training process for the image captioning model is structured around iterative optimization of an encoder–decoder architecture, where learning progresses through multiple complete passes over the dataset, referred to as epochs. Each epoch allows the model to refine its parameters using gradient-based optimization, gradually minimizing the loss between predicted and ground-truth captions.

- **Dataset Size:** The model is trained on a total of **18,542 image–caption pairs**, where each image is associated with multiple Bengali textual descriptions. These pairs provide a rich multimodal dataset that grounds visual content in natural-language semantics.
- **Training Configuration:** Training was conducted for a maximum of **100 epochs**, with **early stopping (patience = 5)** to prevent overfitting. A batch size of $B = 1$ was used, resulting in 18,542 parameter update steps per epoch.

- **Total Iterations:** The total number of training iterations ($\mathcal{I}_{\text{total}}$) is computed as:

$$\mathcal{I}_{\text{total}} = \text{Steps per Epoch} \times \text{Number of Epochs} \tag{3.1}$$

Substituting the dataset-specific parameters yields:

$$\mathcal{I}_{\text{total}} = 18,542 \times 100 = 1,854,200 \tag{3.2}$$

Thus, complete training comprises approximately **1.85 million** update operations before early stopping.

- **Loss Function:** The learning objective minimizes the *sparse categorical cross-entropy* (SCCE) loss over all time steps in each generated caption. Given a target sequence of length T, the loss is defined as:

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T} \log P(y_t \mid y_{1:t-1}, I) \tag{3.3}$$

Here, y_t denotes the ground-truth token at step t, and $P(y_t \mid y_{1:t-1}, I)$ represents the probability assigned by the decoder given previous tokens and the encoded image I. Padding tokens are masked to ensure accurate loss computation.

- **Optimization Strategy:** Parameter updates are performed using the **Adam optimizer**. The training procedure employs **teacher forcing**, where the decoder receives the ground-truth token at each time step, enabling faster convergence and reducing exposure bias during early training.

- **Regularization via Early Stopping:** Validation loss is monitored after every epoch. The model checkpoint corresponding to the lowest validation loss is saved automatically. Training terminates when the validation loss fails to improve for five consecutive epochs, thereby preventing overfitting and ensuring reliable generalization.

- **Training Duration:** All training was executed within the **CodeOcean environment**, with the full experiment completing in approximately **16 minutes and 24 seconds**. Model weights, tokenizer files, and experiment metadata were archived to support reproducibility and efficient re-initialization.

3.5.2 Model Architecture

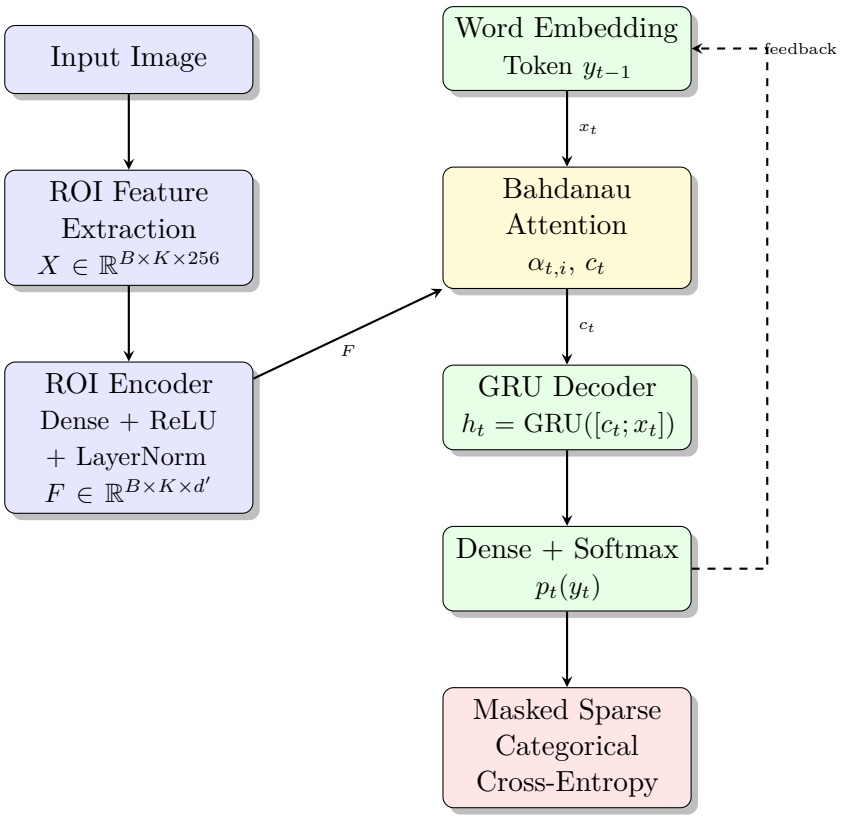
The architecture consists of an encoder–decoder pipeline designed to translate the visual representational space of an image into a coherent Bengali textual description. The architecture has been summarized in Figure 1.

- **Dataset:** The dataset consists of **18,542 training samples**, each containing one image accompanied by multiple Bengali descriptions.

- **Vocabulary Size:** The tokenizer yields a vocabulary of **22,276 unique words**, representing the lexical space the decoder can generate.

- **Caption Length:** The average caption length is approximately **38 tokens**, capturing detailed descriptions of image content.

Figure 1: Encoder–Decoder Architecture with Bahdanau Attention for Bengali Image Captioning



Encoder: ROI Feature Encoder

The encoder maps Region-of-Interest (ROI) feature tensors into a normalized embedding space. Formally:

$$\mathbf{F} = \text{LayerNorm}(\text{ReLU}(W_{\text{enc}} \cdot \mathbf{X} + b_{\text{enc}})) \tag{3.4}$$

where $\mathbf{X} \in \mathbb{R}^{B\times K\times d}$ is the ROI feature tensor (with $d = 256$), and $\mathbf{F} \in \mathbb{R}^{B\times K\times d'}$ represents the transformed embeddings.

Attention Mechanism (Bahdanau Attention)

The decoder integrates Bahdanau attention to dynamically weight image features:

$$e_{t,i} = v^\top \tanh(W_1 f_i + W_2 h_{t-1}) \tag{3.5}$$

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_j \exp(e_{t,j})} \tag{3.6}$$

$$c_t = \sum_i \alpha_{t,i} f_i \tag{3.7}$$

Here, c_t denotes the context vector created by attending to salient visual regions.

Decoder (GRU-based Sequence Generator)

Caption generation proceeds via a gated recurrent unit (GRU):

$$x_t = \text{Embed}(y_{t-1}), \quad z_t = [c_t; x_t] \tag{3.8}$$

$$h_t, o_t = \text{GRU}(z_t, h_{t-1}) \tag{3.9}$$

$$p_t = \text{softmax}(W_o o_t + b_o) \tag{3.10}$$

where p_t is the probability distribution over the next word.

Sparse Categorical Cross-Entropy with Masking

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T} m_t \cdot \log p_t[y_t] \tag{3.11}$$

The mask m_t ensures that padding tokens do not contribute to the loss.

3.5.3 Training Algorithm

The complete training algorithm is summarized in Algorithm 1.

Algorithm 1 Image Captioning Model Training

```
Initialize encoder, decoder, optimizer, checkpoint manager
Set best_val_loss = ∞, patience_counter = 0
for epoch = 1 to EPOCHS do
    Reset training and validation loss trackers
                                                        ▷ — Training —
    for (img_tensor, target) in training_set do
        Initialize decoder hidden state
        Set decoder input = ⟨start⟩
        Encode image: features = Encoder(img_tensor)
        for t = 1 to T do
            Generate predictions via Decoder
            Compute loss for timestep t
            Update dec_input with ground-truth token (teacher forcing)
        end for
        Backpropagate gradients and update parameters
    end for
                                                        ▷ — Validation —
    Compute validation loss
    if validation loss improves then
        Save checkpoint
        Reset patience counter
    else
        Increment patience counter
    end if
    if patience counter reaches limit then
        break
    end if
end for
```

3.6 Discussion

3.6.1 Epoch vs. Loss Behavior

Figure 2 presents the training and validation loss curves for the proposed image captioning model, with early stopping activated at epoch 3. The figure demonstrates several noteworthy learning dynamics.

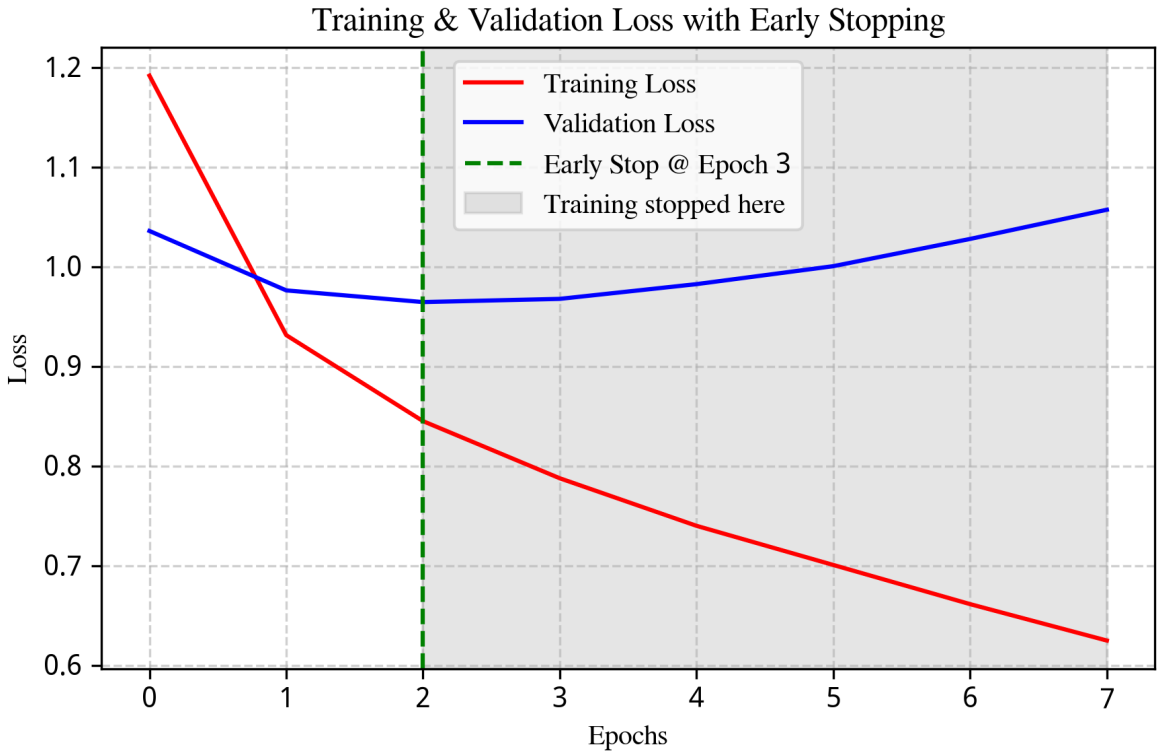


Figure 2: Training and validation loss across epochs for the image captioning model.

The **training loss** exhibits a steady and monotonic decline, dropping from approximately 1.2 at epoch 0 to nearly 0.63 by epoch 7. This trend indicates that the model consistently reduces prediction errors on the training data and successfully learns the underlying captioning patterns.

In contrast, the **validation loss** follows a more nuanced trajectory. After an initial decrease from about 1.04 to nearly 0.97 at epoch 2, the validation loss begins to rise, signaling a divergence between the training and validation performance. This divergence is a classic indicator of **overfitting**, where the model continues to specialize on training patterns but loses generalization capability on unseen data.

The green dashed marker labeled *"Early Stop @ Epoch 3"* highlights the point at which the early stopping criterion was triggered. With a patience value of 5 epochs, the system halted training once the validation loss showed no further improvement. The shaded region beyond epoch 3 indicates the portion of training that was not executed, thereby preventing the model from overfitting further.

Importantly, the early stopping mechanism preserved the checkpoint corresponding to the **optimal validation loss at epoch 2**, thus ensuring that the model used for evaluation reflects the best possible generalization performance. This strategic termination effectively balances minimizing training error while maintaining robust performance on unseen image–caption pairs.

3.6.2 Model Evaluation Results

The performance of the trained model was assessed on a held-out test set of seven images using the **BLEU (Bilingual Evaluation Understudy)** metric as the primary measure for caption quality. BLEU evaluates the n-gram overlap between model-generated captions and reference captions, producing a score between 0 and 1, where higher values

denote closer correspondence to human-written descriptions.

Quantitative Evaluation

The quantitative results reveal a diverse performance profile across the test samples. The model achieved a **mean BLEU score of 0.5157**, suggesting a moderate alignment with ground-truth captions. The **median BLEU score of 0.4609** further indicates that at least half of the generated captions fall below this threshold, emphasizing variability in caption quality.

BLEU scores ranged from **0.2212 (minimum)** to **0.7071 (maximum)**, yielding a performance spread of approximately **0.4859**. This appreciable spread reflects substantial inconsistency across image types and content complexity. Additionally, the **standard deviation of 0.1803** reinforces this variability and implies that while the model excels on certain images, it performs poorly on others.

Overall, the quantitative metrics highlight that although the model demonstrates the ability to generate relevant captions in many cases, its performance is not uniformly reliable across different visual contexts.

Qualitative Analysis

A closer qualitative inspection of generated captions uncovers distinct behavioral patterns. On higher-performing examples (BLEU 0.7071), particularly for images containing **groups of people**, **pets**, or **simple activities**, the model correctly identifies major entities and actions. Even though syntactic imperfections occasionally arise, the captions preserve core semantic content and demonstrate meaningful scene comprehension.

However, in lower-scoring cases (BLEU between 0.2212 and 0.4609), several issues become prominent:

- **Semantic drift:** The model sometimes produces captions that stray from the true content, mislabeling objects or misinterpreting the scene context.
- **Incorrect object identification:** Outdoor/indoor settings are frequently misclassified, and unrelated or nonexistent entities are occasionally introduced.
- **Presence of unknown tokens (`<unk>`):** This suggests vocabulary limitations, likely caused by insufficient lexical coverage in the training corpus.
- **Partial relevance:** Some captions contain accurate elements but fail to capture critical visual details, reducing semantic completeness.

These observations indicate that while the model captures broad thematic cues, it often lacks the granularity needed for fine-grained caption accuracy.

4 Results and Analysis

4.1 Overview

This chapter summarizes training and evaluation of the image-captioning models.

- **CNN (EfficientNetB4)** — transformer-based encoder with an EfficientNet-B4 backbone and a Transformer decoder.
- **CNN (InceptionV3)** — an encoder-decoder model inspired by "Show, Attend and Tell" that uses an InceptionV3 encoder, Bahdanau attention and an RNN (GRU) decoder.

Below we summarize the quantitative training traces, highlight qualitative outputs, and discuss differences between the two model families together with recommended next steps.

4.2 EfficientNet-B4 encoder + Transformer decoder

The EfficientNet-B4 experiment was trained for up to 50 epochs with the final saved training artifact showing 15 completed epochs (early stopping / checkpointing triggered). Key hyperparameters include image input size $(380, 380)$, sequence length 8, embedding dimension 768, batch size 64 and a learning-rate schedule with warmup followed by a base LR of 1×10^{-4}.

Table 4.1 reports selected training metrics from the final epoch (epoch 15): training loss decreased steadily from an initial high value (8.01 at epoch 1) to 3.19 at epoch 15, while training token-level accuracy rose from 0.01 to 0.40. Validation loss also decreased from an initial very large value (36.09) down into the low twenties, and validation accuracy reached 0.27 by epoch 15. The training run took approximately 7 minutes (wall-clock).

Metric	First epoch	Final recorded epoch (15)
Training loss	8.0074	3.1921
Training acc	0.0131	0.4014
Validation loss	36.0912	22.2223
Validation acc	0.0738	0.2739

Table 4.1: Training metrics for the EfficientNet-B4 model

Interpretation: the EfficientNet-based Transformer model shows a strong downward trend in training loss and steadily improving token-level accuracy. The validation loss is substantially higher than the training loss across most epochs, which suggests either a distributional difference between the training and validation splits (or evaluation mismatch) or that further regularization (or more training) is required.

4.3 InceptionV3 encoder + Bahdanau attention + GRU decoder

The InceptionV3 experiment implements the classic Show–Attend–and–Tell style encoder–decoder with additive (Bahdanau) attention. The training artifacts include precomputed image features (cached .npy files), a tokenizer, and checkpoints.

The loss values show steady training loss improvement from 1.20 (epoch 0) down to 0.66 (epoch 7). Validation loss remains near 1.0 throughout the run and slightly increases toward later epochs (0.98 Ø 1.05), indicating a smaller but persistent generalization gap compared with training loss.

Metric	First epoch	Final recorded epoch (7)
Training loss	1.20	0.66
Validation loss	1.05	1.05

Table 4.2: Training metrics for the InceptionV3 model

Interpretation: The InceptionV3-based model attains much lower numerical loss values than the EfficientNet model. This numeric difference should not be interpreted as an absolute comparison of model quality because the two experiments use different decoder architectures and loss formulations, vocabulary sizes and decoding strategies (greedy vs stochastic sampling). Instead, focus on within-experiment trends: InceptionV3 shows consistent training improvement and a relatively stable validation loss, suggesting reasonable generalization for the trained regime.

4.4 Qualitative outputs and attention

Both models have implementation differences, and these differences affect qualitative behavior:

- The EfficientNet model uses greedy decoding (argmax) during generation by default; captions are therefore deterministic given the same checkpoint.
- The InceptionV3 model uses stochastic sampling (‘tf.random.categorical‘) at inference, producing diverse captions for the same image; for repeatable evaluation use the deterministic sampling or beam search.

Attention visualizations produced by the InceptionV3 pipeline are particularly useful to verify that the model focuses on sensible spatial regions while decoding.

4.5 Comparative analysis and caveats

Direct numeric comparison between the two runs is not advisable without normalizing the evaluation protocol; below we expand that caveat and summarize the observed differences in five concrete dimensions in Table 4.3.

4.5.1 Convergence

The EfficientNet-B4 + Transformer run shows a strong drop in training loss from 8.01 to 3.19 over the recorded epochs, with token-level training accuracy rising to 0.40; however its validation loss remains comparatively high (final recorded value 22.22). This pattern indicates fast in-run optimization on the training objective but a persistent training–validation gap.

The InceptionV3 + Bahdanau/GRU run attains much lower numerical losses (training loss 0.66 at the last logged epoch and validation loss 1.05). Within its own training regime the Inception experiment exhibits faster apparent convergence to a lower-loss plateau, but note the two runs use different loss scalings and vocabulary/tokenization which prevent direct numerical comparison of loss magnitudes.

4.5.2 Stability

Stability refers to how smoothly metrics evolve across epochs and how sensitive the run is to overfitting. The EfficientNet Transformer shows steady training improvement but a larger and more variable validation loss, suggesting either unstable generalization or an evaluation/preprocessing mismatch between training and validation. The InceptionV3 run shows steadier validation loss (a near-flat curve around 1.0) with smaller per-epoch validation fluctuations, indicating higher apparent validation stability under the current experimental settings.

4.5.3 Interpretability

Interpretability differs by architecture and by the visualization tools implemented: the InceptionV3 pipeline explicitly produces spatial attention visualizations (saved under ‘generated_images/‘) coming from the Bahdanau additive attention—these visualizations map decoded tokens to image regions and are directly useful for human inspection of model focus. The EfficientNet Transformer uses token/self-attention in the decoder; unless cross-attention visualizations were implemented, its attention is less directly mapped to image pixels in the current outputs. In short: the InceptionV3 artifacts provide immediate spatial interpretability; the EfficientNet run may require additional tooling (cross-attention heatmaps or gradient-based saliency) to reach the same level of spatial explainability.

4.5.4 Decoding behavior and its effects

Decoding choices materially affect evaluation and qualitative captions. The EfficientNet experiment was decoded deterministically (greedy/argmax) in the saved outputs, producing repeatable captions; the InceptionV3 experiment used stochastic sampling (‘tf.random.categorical‘) during generation, yielding diverse captions for the same image. As a result, BLEU/CIDEr and other n-gram metrics computed from the two runs are not directly comparable unless the same deterministic decoding (greedy or beam search with fixed seeds) is used for both. Stochastic decoding also inflates apparent diversity (helpful for qualitative analysis) but reduces metric stability for single-run comparisons.

4.5.5 Limitations of Comparison

The following limitations should be stated explicitly in any comparative claim; this language is suitable for an IEEE-style ”Limitations” subsection and can be used verbatim in the paper.

Experimental and implementation differences

The two experiments differ in multiple orthogonal dimensions (encoder backbone, decoder family, attention implementation, tokenizer/vocabulary, loss scaling and 'from_logits' settings, decoding strategy and evaluation scripts). Any direct numeric comparison (for example, of cross-entropy loss values or raw BLEU numbers obtained under different decoding policies) is therefore confounded by these implementation differences.

Evaluation protocol and metric sensitivity

Evaluation metrics such as BLEU and CIDEr are sensitive to decoding determinism, vocabulary choices and tokenization. Stochastic decoding increases output variance and reduces the interpretability of single-run metric snapshots. Without a unified decoding and tokenization protocol, numeric metrics can be misleading.

Dataset splits and preprocessing

If training/validation/test preprocessing differs (image resizing, normalization, or text normalization/token filtering), observed validation gaps may reflect preprocessing mismatches rather than model capacity. We did not re-run training under a single unified pipeline; the artifacts are analyzed as-is.

Suggested mitigation

To support reproducible and fair comparisons, re-run evaluation under a single, shared protocol: (1) use the same tokenizer and vocabulary; (2) decode deterministically (beam search or greedy) with fixed seeds when computing metrics; (3) align preprocessing (image resize/normalization) between train and validation; and (4) compute metrics over multiple seeds or multiple samples for stochastic decoders.

4.6 Summary

Both the EfficientNet-B4 Transformer and the InceptionV3 + Bahdanau attention experiments show meaningful training progress and produce usable caption outputs. The numerical training traces differ in scale and trend because of architectural and implementation differences; the next step is to standardize the evaluation protocol (tokenizer, decoding, metric computation) and then compare BLEU/CIDEr on the same held-out test set. Qualitatively, attention visualizations from the InceptionV3 experiment are already useful for interpreting model focus; the EfficientNet transformer (with its larger embedding size) is a promising candidate for further tuning and checkpoint selection based on BLEU.

Dimension	EfficientNet-B4 + Transformer	InceptionV3 + Bahdanau/GRU
Convergence	Training loss drops strongly (8.01 → 3.19) with token accuracy reaching ∼0.40; validation loss remains high (∼22.22), indicating pronounced train–val gap.	Lower numerical training loss (1.20 → 0.66) with stable validation around ∼1.05; converges to lower-loss plateau within its regime.
Stability	Training metrics improve steadily but validation shows larger fluctuations and higher variance, suggesting possible overfitting or evaluation mismatch.	Validation curve is comparatively flat and less noisy (∼1.0), indicating higher apparent stability under current settings.
Interpretability	Decoder uses token/self-attention (transformer); cross-attention visualizations not provided by default—spatial explainability limited without extra tooling.	Produces spatial attention visualizations (Bahdanau) mapping tokens to image regions; immediately useful for human inspection.
Decoding Behavior	Deterministic (greedy/argmax) decoding for saved outputs—repeatable but may yield terse captions; metrics stable for single-run evaluation.	Stochastic sampling (`tf.random.categorical`) for generation—yields diverse captions but increases metric variance and complicates comparisons.
Error Typology	Large train–val gap suggests lexical/tokenization mismatches, preprocessing issues, and overfitting; hallucinations and attribute errors expected.	Hallucinations and misattributed objects occur but easier to diagnose via attention maps; stochastic decoding can introduce repetition/truncation artifacts.
Recommended Follow-up	Verify tokenizer & preprocessing alignment; add cross-attention or saliency visualizations; apply targeted regularization and small-scale finetuning.	Recompute metrics under deterministic decoding (or aggregate samples); use attention maps for targeted error analysis and qualitative illustrations.

Table 4.3: Qualitative evaluation: concise comparison of convergence, stability, interpretability, decoding behavior, and error typology between the two experiments.

5 Challenges, Future Works, and Conclusion

5.1 Challenges

The development of the Bengali image captioning system presented in this study was accompanied by a wide range of challenges spanning dataset limitations, computational constraints, modeling complexity, and language-specific barriers. These challenges directly shaped the performance, generalization capacity, and robustness of the model.

5.1.1 Dataset Size and Memory Constraints

- **Large Dataset:** The dataset contains 18,542 image–caption pairs, which, although moderate compared to large-scale English datasets such as MS-COCO, introduces significant demands on storage, preprocessing, and efficient data loading pipelines.
- **Memory Limitations:** Due to restricted hardware resources, it was not feasible to load all extracted image embeddings and tokenized captions simultaneously into memory. This necessitated the use of data generators, batch-wise loading, and on-the-fly preprocessing to ensure stable and memory-efficient training.

5.1.2 Model Complexity and Training Time

- **Computationally Intensive Architecture:** The CNN–LSTM pipeline, while conceptually effective, is computationally expensive. Feature extraction using convolutional networks combined with long-sequence modeling in LSTMs significantly increases processing overhead.
- **Extended Training Duration:** Training over tens of thousands of samples requires numerous epochs, each involving extensive backpropagation. Limited GPU capacity further prolonged experimentation cycles, reducing opportunities for iterative fine-tuning.

5.1.3 Hyperparameter Tuning and Optimization

- **Numerous Hyperparameters:** Optimal choices for learning rate, batch size, embedding dimension, LSTM units, dropout rate, and optimizer settings strongly influence model outcomes. Finding a stable configuration required extensive trial-and-error experimentation.
- **Optimization Difficulties:** Stabilizing the training process posed persistent challenges. With a relatively small dataset and a highly morphological language, the model was prone to both overfitting and underfitting, complicating convergence.

5.1.4 Caption Quality and Model Performance

- **Limited Caption Quality Improvements:** Despite multiple training attempts, regularization strategies, and early stopping, BLEU score improvements plateaued early. High variance across samples suggests inadequate generalization.
- **Semantic and Grammatical Errors:** Generated captions often contained incorrect word order, missing context, or incomplete phrases. The complex morphology and syntactic flexibility of Bengali contributed significantly to these inconsistencies.
- **Object Misidentification:** The model struggled with fine-grained object detection and spatial reasoning, leading to captions that were semantically vague or partially incorrect.

5.1.5 Dataset Quality and Quantity Issues

- **Limited Dataset Diversity:** While containing over 18K samples, the dataset lacked the breadth and richness of large multimodal corpora. Many rare or complex scene types were underrepresented.
- **Inconsistent Caption Style:** Variations in caption structure, grammatical accuracy, and descriptive level introduced noise, making it difficult for the model to learn a stable language pattern.
- **Vocabulary Coverage Constraints:** The frequent appearance of unknown tokens (`<unk>`) revealed limitations in vocabulary coverage, restricting the model's descriptive capacity.

5.1.6 Training Environment Limitations

- **Restricted GPU Resources:** Training was carried out under limited computing capacity, constraining batch sizes, training duration, and exploration of advanced architectures.
- **Slower Experimentation Cycles:** Limited resources slowed hyperparameter tuning, architecture exploration, and ablation studies.

5.1.7 Dataset and Result Bias

- **Distributional Bias:** Overrepresentation of particular image categories (e.g., people, animals) biased the model towards generating captions dominated by common scene elements.
- **Caption Style Bias:** If training captions emphasized high-level semantics rather than details, the model reproduced this tendency.
- **Evaluation Bias:** BLEU's reliance on n-gram overlap may underestimate semantically valid captions with different phrasing.

5.1.8 Language-Specific Challenges

- **Morphological Complexity of Bengali:** Bengali's rich inflection, compound formation, and flexible syntax make modeling substantially harder than captioning tasks in English.
- **Lack of Bengali Benchmarks:** Unlike English, Bengali lacks large-scale datasets and pretrained captioning models, reducing opportunities for transfer learning.

5.2 Future Work

Several research directions emerge from the limitations and insights identified in this study. Advancing these areas may lead to more accurate, fluent, and deployable Bengali image captioning systems capable of serving diverse real-world applications.

5.2.1 Advancing Model Architectures

Transformer-based visual encoders such as ViT and Swin Transformer have shown strong performance in recent vision tasks [12] and may substantially outperform traditional CNN–LSTM architectures [9, 17]. Cascaded or modular neural networks, which have demonstrated effectiveness in low-resource environments [13], represent another promising direction for structured feature extraction and caption refinement. Additionally, ensemble-based approaches combining multiple CNN or transformer backbones [26] may yield richer visual embeddings and reduce uncertainty in object recognition.

5.2.2 Improving Attention, Alignment, and Localization

Future work may focus on enhanced attention mechanisms, including multi-head or hierarchical attention, to improve the model's ability to highlight salient regions during caption generation. Incorporating region-based feature extraction techniques—such as leveraging object detection models like Faster R-CNN—could further improve spatial awareness and fine-grained caption quality.

5.2.3 Dataset Expansion and Annotation Improvements

Expanding the dataset with broader visual and linguistic diversity is essential for improving generalization [8]. Enhancing annotation quality through greater consistency in grammar and descriptive detail would strengthen the learning signal. Multiple captions per image should be included to increase linguistic diversity and reduce evaluation variance, while category rebalancing may mitigate dataset-induced biases.

5.2.4 Leveraging Pretraining and Transfer Learning

Pretrained visual encoders such as EfficientNet have proven effective in Bengali NLP and vision tasks [10] and may further improve feature representation when fine-tuned. Multilingual or Bengali-specific language models including mBERT, XLM-R, or Bengali transformers [25] can contribute to better morphological and grammatical handling. Additionally, contrastive pretraining approaches similar to CLIP may strengthen multimodal alignment for Bengali image–text pairs.

5.2.5 Improving Training Efficiency

Memory-efficient training strategies—such as mixed precision, gradient checkpointing, or tensor offloading—may enable larger models to be trained on limited hardware. Distributed training using cloud-based GPUs or TPUs can accelerate large-scale experimentation. Furthermore, model compression techniques including pruning and quantization can support deployment on mobile or resource-constrained devices.

5.2.6 Mitigating Dataset and Model Bias

Future efforts should incorporate explicit bias evaluation metrics to measure linguistic and visual disparities within the dataset. Balanced sampling strategies or class-balanced loss functions may reduce overfitting toward dominant object categories and improve fairness across diverse image types.

5.2.7 Enhancing Caption Fluency

Improving caption fluency may involve grammar-aware decoding, morphological rule integration, or hybrid neural–rule-based approaches. Reinforcement learning techniques using semantic-oriented metrics such as CIDEr, SPICE, or human feedback as rewards may further improve the semantic fidelity of generated captions.

5.2.8 Developing APIs for Organizational Integration

A significant future direction is the development of production-ready, scalable APIs that allow organizations to integrate Bengali image captioning into their operational workflows. Such APIs would enable automated caption generation without requiring expertise in machine learning. The framework should also support domain-specific fine-tuning using an organization's proprietary dataset, facilitating adaptation to specialized domains such as healthcare, agriculture, media archives, or e-commerce. A modular backend supporting customizable encoders, tokenizer updates, and vocabulary expansion will further enhance usability. Finally, security mechanisms, model versioning, and monitoring capabilities will be essential for reliable large-scale deployment.

5.3 Conclusion

This study explored the development of a Bengali image captioning model that integrates convolutional neural networks for feature extraction and recurrent neural networks for natural language generation. Despite operating within a low-resource environment, the system successfully produced meaningful, human-readable captions and demonstrated the feasibility of multimodal learning for underrepresented languages such as Bengali.

The challenges encountered—ranging from dataset limitations and computational constraints to linguistic complexity and evaluation biases—highlight the demanding nature of Bengali caption generation. Moderate BLEU scores and variations in caption quality reflect the combined impact of dataset imbalance, limited training resources, and the absence of large-scale Bengali benchmarks.

However, this research provides a strong foundation for continued advancement. The findings underscore the need for stronger architectures, richer datasets, improved annotation quality, and deeper attention mechanisms. The future directions identified in this

chapter outline promising pathways toward more accurate, detailed, and linguistically fluent captioning systems.

Ultimately, this work contributes meaningfully to the growing field of Bengali AI research by illuminating the challenges, demonstrating practical solutions, and establishing a basis upon which future multimodal systems can be built. By expanding datasets, refining architectures, and integrating multilingual pretrained models, future researchers can move toward highly expressive, contextually aware Bengali captioning systems that support accessibility, information retrieval, and enriched digital communication.

Bibliography

[1] Al Faraby, H., et al. Image to bengali caption generation using deep cnn and bidirectional gated recurrent unit. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)* (2020), IEEE, pp. 27–34.

[2] Bhattacharjee, A., Hasan, T., Ahmad, W., Mubasshir, K. S., Islam, M. S., Iqbal, A., Rahman, M. S., and Shahriyar, R. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022* (Seattle, United States, July 2022), Association for Computational Linguistics, pp. 1318–1327.

[3] Bhattacharjee, A., Hasan, T., Ahmad, W., and Shahriyar, R. Banglanlg and banglat5: Benchmarks and resources for evaluating low-resource natural language generation in bangla. In *Findings of the Association for Computational Linguistics: EACL 2023* (2023), pp. 714–723.

[4] Bhattacharjee, A., Hasan, T., Ahmad, W. U., and Shahriyar, R. Banglanlg: Benchmarks and resources for evaluating low-resource natural language generation in bangla. *arXiv preprint arXiv:2205.11081* (2022).

[5] Bhuiyan, A., Hossain, E., Hoque, M. M., and Dewan, M. A. A. Enhancing image caption generation through context-aware attention mechanism. *Heliyon 10*, 17 (2024).

[6] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).

[7] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online, July 2020), Association for Computational Linguistics, pp. 8440–8451.

[8] Faiyaz Khan, M., Sadiq-Ur-Rahman, S. M., and Saiful Islam, M. Improved bengali image captioning via deep convolutional neural network based encoder-decoder model. *Algorithms for Intelligent Systems* (2021), 217–229.

[9] Geetha, G., Kirthigadevi, T., Ponsam, G. G., Karthik, T., and Safa, M. Image captioning using deep convolutional neural networks (cnns). *Journal of Physics: Conference Series 1712* (12 2020), 012015.

[10] Hasan, M., Alfaz, N., Alam, M., Rahman, A., Shakhawat, H., and Rahman, S. Detection of Parkinson's disease from T2-weighted magnetic resonance imaging scans using EfficientNet-V2. In *2023 26th International Conference On Computer And Information Technology (ICCIT)* (2023), pp. 1–6.

[11] Hochreiter, S., and Schmidhuber, J. Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780.

[12] Hossain, M., Shahriar, G., Syeed, M., Uddin, M., Hasan, M., Shivam, S., and Advani, S. Region of interest (ROI) selection using vision transformer for automatic analysis using whole slide images. *Scientific Reports 13* (2023), 11314.

[13] Islam, M., Rahman, W., Hossain, M., Roksana, K., Azpíroz, I., Diaz, R., Ashraf, I., and Samad, M. Medicinal plant classification using particle swarm optimized cascaded network. *IEEE Access 12* (2024), 42465–42478.

[14] Kabir, M. K., et al. Automatic bengali image captioning using efficientnet-transformer network. In *2023 22nd International Symposium on Communications and Information Technologies (ISCIT)* (2023), IEEE, pp. 110–115.

[15] Kamal, A. H., Jishan, M. A., and Mansoor, N. Textmage: The automated bangla caption generator based on deep learning. In *2020 International Conference on Decision Aid Sciences and Application (DASA)* (2020), IEEE, pp. 1–6.

[16] Kim, Y. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha, Qatar, Oct. 2014), Association for Computational Linguistics, pp. 1746–1751.

[17] Liu, S., Bai, L., Hu, Y., and Wang, H. Image captioning based on deep neural networks. *MATEC Web of Conferences 232* (2018), 01052.

[18] Lundberg, S. M., and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)* (2017).

[19] Mohan, R. K., Sureshkumar, S., and Sivasubramaniam, V. Explainable image captioning using cnn-cnn architecture and hierarchical attention. *arXiv preprint arXiv:2407.09556* (2024).

[20] Natarajan, P., and Nambiar, A. Vale: A multimodal visual and language explanation framework for image classifiers using explainable ai and language models. *arXiv preprint arXiv:2408.12808* (2024).

[21] Palash, M. A. H., et al. Bangla image caption generation through cnn-transformer based encoder-decoder network. In *Proceedings of International Conference on Fourth Industrial Revolution and Beyond 2021* (2022), Springer Nature Singapore, pp. 123–130.

[22] Rahman, M., et al. Chittron: An automatic bangla image captioning system. *Procedia Computer Science 154* (2019), 636–642.

[23] Ribeiro, M. T., Singh, S., and Guestrin, C. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016).

[24] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017).

[25] Shah, F. M., Humaira, M., Jim, Ami, A. S., and Paul, S. Bornon: Bengali Image Captioning with Transformer-based Deep learning approach. *arXiv* (09 2021).

[26] Sifat, S., Hossain, M., Tonny, S., Majumder, B., Mahajabin, R., and Shakhawat, H. Android ransomware attacks detection with optimized ensemble learning. In *International Conference On Cybersecurity, Cybercrimes, And Smart Emerging Technologies* (2022), pp. 41–53.

[27] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems 30* (2017).

[28] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015).

[29] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)* (2015).

[30] Yu, J., et al. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917* (2022).