Heliyon 10 (2024) e36272

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Enhancing image caption generation through context-aware attention mechanism

Ahatesham Bhuiyan [a], Eftekhar Hossain [a], Mohammed Moshiul Hoque [b,*], M. Ali Akber Dewan [c]

[a] Department of Electronics and Telecommunication Engineering, Chittagong University of Engineering and Technology, Chittagong, 4349, Bangladesh

[b] Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chittagong, 4349, Bangladesh

[c] School of Computing and Information Systems, Faculty of Science and Technology, Athabasca University, Athabasca, AB T9S 3A3, Canada

ARTICLE INFO

Keywords:
Image captioning
Encoder-decoder
Computer vision
Cross-domain transfer
Natural language processing
Attention mechanism

ABSTRACT

Image captioning, the process of generating natural language descriptions based on image content, has garnered attention in AI research for its implications in scene understanding and human-computer interaction. While much prior research has focused on caption generation for English, addressing low-resource languages like Bengali presents challenges, particularly in producing coherent captions linking visual objects with corresponding words. This paper proposes a context-aware attention mechanism over semantic attention to accurately diagnose objects for image captioning in Bengali. The proposed architecture consists of an encoder and a decoder block. We chose ResNet-50 over the other pre-trained models for encoding the image features due to its ability to solve the vanishing gradient problem and recognize complex object features. For decoding generated captions, a bidirectional Gated Recurrent Unit (GRU) architecture combined with an attention mechanism captures contextual dependencies in both directions, resulting in more accurate captions. The paper also highlights the challenge of transferring knowledge between domains, especially with culturally specific images. Evaluation of three Bengali benchmark datasets, namely *BAN-Cap*, *BanglaLekhaImageCaption*, and *Bornon*, demonstrates significant performance improvement in METEOR score over existing methods by approximately 30%, 18%, and 45%, respectively. The proposed context-aware, attention-based image captioning system significantly outperforms current state-of-the-art models in Bengali caption generation despite limitations in reference captions on certain datasets.

1. Introduction

Automatic image captioning (AIC) refers to a computer vision technique that employs machine learning algorithms to generate descriptive textual explanations of the content depicted in an image. It combines image recognition and natural language processing techniques to enable machines to comprehend visual information and provide human-readable captions, enhancing the accessibility and understanding of visual content [1]. Automated image captioning has emerged as a vital area of research in deep learning, with

[*] Corresponding author.
E-mail addresses: u1708002@student.cuet.ac.bd (A. Bhuiyan), eftekhar.hossain@cuet.ac.bd (E. Hossain), moshiul_240@cuet.ac.bd (M.M. Hoque), adewan@athabascau.ca (M. Ali Akber Dewan).

https://doi.org/10.1016/j.heliyon.2024.e36272
Received 12 April 2024; Received in revised form 7 August 2024; Accepted 13 August 2024

Available online 19 August 2024

numerous practical applications. One such application is in the image search field, which can yield better results and enhance the user experience. It can also play a crucial role in early childhood education by fostering children's interest in and understanding visual content. Additionally, image captioning can benefit blind individuals by converting images into speech, enabling them to access and comprehend digital media. Beyond these, it has the potential to aid in retrieving information via concept-based indexing [2], automate self-driving cars [3], and even caption surveillance footage [4] in real-time to prevent theft, crime, or accidents, and ensure public safety.

While there has been significant progress in image captioning research concerning high-resource languages, more research in Bengali needs to be done. Therefore, this work aims to explore and contribute to automatic image caption generation in Bengali. Automatic image captioning in a resource-constrained language (e.g., Bengali) is critical due to the need for sizable and high-quality image caption datasets. Moreover, the complex morphological structures, scarcity of language processing tools, and lack of relevant computational linguistic resources made the task more challenging. Existing datasets like Flicker8k [5] and MSCOCO [6] are biased towards Western culture, leading to inaccurate results while a caption is written in Bengali. For instance, a young boy wearing the typical male attire of Bangladesh, the lungi, may be misidentified as a female due to the model's cultural bias [7]. The availability of only two superior public datasets in Bengali, Bornon [8] and BanglaLekhaImageCaptions [9], is also a significant obstacle in the research. Therefore, extensive efforts should emerge to develop a reliable and effective method for generating image captions in Bengali.

Image captioning is a challenging task that relies on an encoder-decoder framework, with a CNN-RNN-based architecture [10,11] being a popular choice. However, existing methods overlook spatial relationships between objects in the image, which is essential for human reasoning and understanding the physical world. Therefore, exploring new approaches that incorporate this information is crucial. Additionally, developing a model that can generate coherent captions in the Bengali language requires considering the language's unique features and nuances. Although recent studies have focused on Bengali image captioning, the majority of them use a CNN-RNN-based architecture [9,12–14] where they employed InceptionV3, VGG16 and Xception techniques for image feature extraction. Only a few consider attention-based models [15,16] whose performance needs to be more reliable. Moreover, cross-domain transfer approaches for image captioning tasks in Bengali need to be researched. To address the drawbacks above, we investigate the following two research questions throughout this work.

- **RQ1:** How do context-aware representations of image features beat the conventional encoder-decoder architectures' performance?
- **RQ2:** Can cross-domain transfer effectively generates out-of-domain image captions accurately?

This research aims to demonstrate sub-region-based object detection within image captioning, achieved through context-aware attention-based encoder-decoder architecture. We propose a sequence to the sequence model where ResNet50 has been used as an encoder and attention with GRU for decoding the caption for the encoded image. The choice of the ResNet-50 architecture was motivated by its ability to capture features ranging from low-level details to high-level semantic content, making it ideal for understanding image context and addressing vanishing gradient issues. On the other hand, context-aware attention with GRU enriches caption quality by selectively emphasizing relevant image regions. More specifically, in this work, the architectural choice has been made based on several key factors. Firstly, we were motivated to choose ResNet and BiGRU based on previous studies of image captioning in other languages (e.g., English and Hindi). Secondly, the Bengali language comprises rich inflections and derivations. This complexity requires a model that can handle intricate morphological variations effectively. According to some studies [17], bidirectional GRU (BiGRU) is well-suited because it captures dependencies in both directions, which is crucial for understanding and generating morphologically complex sentences. Third, Bengali captions may include objects not commonly found in Western datasets. The use of ResNet-50 helps in capturing detailed and complex features from images, which is essential for recognizing culturally specific items accurately. Lastly, Bengali often relies heavily on context for meaning, especially due to its free word order compared to English. The context-aware attention mechanism enhances the model's ability to focus on relevant parts of the image that are contextually significant for generating accurate captions. This helps in aligning visual features with the appropriate linguistic constructs in Bengali. Our proposed model outperformed existing models regarding three publicly available datasets: BAN-Cap [18], BanglaLekhaImageCaptions [9], and Bornon [8] in terms of several standard evaluation metrics such as BLEU and Meteor. Besides, the incorporation of context-aware attention has (i) improved the descriptive accuracy of captions as it considers the relationships between objects in an image and (ii) generated captions that are more contextually coherent. The critical contributions to this research study can be attributed to the following answers to the research questions (ARQ1 and ARQ2):

- **ARQ1:** This work demonstrates the superiority of incorporating ResNet-50 and a context-aware attention network over the conventional encoder-decoder architecture in Bangla image captioning. The proposed approach outperforms existing methods by leveraging contextual information and achieving higher accuracy in generating captions.
- **ARQ2:** Through extensive experiments on three benchmark datasets, we analyze the adaptiveness of image captioning outcomes across domains, showcasing the potential of cross-domain transfer for generating accurate out-of-domain image captions.

The rest of the article is structured as follows. Section 2 provides an overview of related research on image captioning. Section 3 explains the architecture of the proposed method, followed by details of the experimental setup, key findings, and qualitative analysis in Section 4. Finally, Section 6 concludes the work with an outline of a few scopes for further improvements.

2. Related work

This segment presents pertinent context regarding previous research on generating image captions and attention. Numerous approaches have recently been proposed for producing descriptions of images.

2.1. Encoder-decoder approach

Earlier image captioning approaches are predominantly rule-based and generate slotted captions using object detection techniques [19,20]. Recently, several methods [21,10,22] employed encoder-decoder architecture motivated by a sequence-to-sequence learning approach. In the encoder portion, a pre-trained CNN extracts image features. In contrast, Recurrent Neural Networks (RNNs) are used in the decoder part to convert this representation into a natural language description [11]. The encoder-decoder framework [23] of machine translation is highly appropriate, as it bears a resemblance to the process of "translating" an image into a sentence.

2.2. Attention and transformer based approach

In the encoder-decoder approach, the encoder may not effectively capture spatial information and focus on important image regions. It treats the entire image equally, which can be suboptimal for image captioning, where different regions may contribute differently to the description. When using recurrent neural networks (RNNs) as decoders, long sequences of words can lead to vanishing gradient problems during training. Therefore, Bahdanau et al. [24] employed the attention mechanism to concentrate on specific portions of the input sentence. This lets the decoder possess more pertinent information during a particular time step. The foundations of the work conducted in Image Captioning models, as proposed by Vinyals et al. [11] and Xu et al. [25], respectively. Xu et al. [25] introduced the initial model of visual attention for generating captions for images, providing a selection between "hard" pooling, which focuses on the most probable region of attention, and "soft" pooling, which combines spatial features with attention weights. Chen et al. [6] further expanded on this concept by incorporating spatial and channel-wise attention within a convolutional neural network. In contrast, You et al. [26] proposed an alternative approach that involves the identification of significant semantic attributes from the image and employed bottom-up strategies to extract important features. Li et al. [27] proposed a method for a transformer-style structure, called Comprehending and Ordering Semantics Networks (COS-Net), that unifies an enriched semantic comprehension and learnable semantic ordering processes into a single architecture for image captioning. Empirical evidence shows that COS-Net surpasses the state-of-the-art approaches on COCO [28] and achieves the best CIDEr score of 141.1%. However, this approach may have limitations in terms of scalability and generalization to other datasets or domains. CIDEr (Consensus-based Image Description Evaluation) [29] is a widely employed metric for appraising the caliber of image explanations engendered by automatic image captioning systems. It measures how well the generated captions align with human-generated reference captions. A CIDEr score surpassing 100% inherently denotes that the produced captions are notably superior to consensus and diversity compared to the reference captions within the specified assessment dataset. It quantitatively measures enhancements in the caliber of captions accomplished by models for image captioning. Fang et al. [30] introduced a new image captioning model called ViT-CAP, using a vision transformer-based approach without relying on a separate object detector to extract regional features. The model achieved 138.1 CIDEr scores on COCO-caption Karpathy-split [28], 93.8 and 108.6 CIDEr scores on Nocaps [31] and Google-CC captioning datasets [32], respectively. Luo et al. [33] proposed a Semantic-Conditional Diffusion Networks (SCD-Net) based approach for image captioning. This work breaks the conventions of learning transformer-based encoder-decoders. It uses a diffusion model to capture the dependency among discrete words and pursue complex visual-language alignment in image captioning. The results show that SCD-Net consistently outperforms state-of-the-art non-autoregressive approaches and achieves comparable or better performance on the COCO dataset [28].

2.3. Image captioning in bangla

Few studies focused on image captioning in Bengali. Rahman et al. [9] introduced the Bengali image caption dataset called *BanglaLekhaImageCaption*, consisting of only two captions per image. They used VGG16 to extract image features and stacked LSTM layers to generate captions. Their method achieved a lower BLEU score of 0.025 without considering the corpus-level word information. Deb et al. [14] proposed a CNN-LSTM-based approach for Bengali image captioning using the bilingual dataset (*Flickr8k-BN*). They employed a FastText word embedding model to train two baseline multimodal models (Par-Inject and Merge), evaluated with BLEU, METEOR, CIDEr, and ROUGE metrics. Although the merged architecture yielded the best BLEU-4 score of 0.22, the models' limitations were their reliance on culturally biased Flickr8k images and generated captions that were not entirely fluent in Bengali. Kamal et al. [13] introduced a CNN-LSTM-based encoder-decoder architecture and achieved the highest BLUE-1 score of 0.667 on the BanglaLekha dataset. Khan et al. [34] presented a multimodal image captioning system that utilizes a pre-trained ResNet50 model for visual feature extraction and a one-dimensional CNN for encoding caption sequence information. The authors achieved the highest BLEU-1 score of 0.651. However, the 1D-CNN approach may miss important details and affect the generation of fluent captions. Jishan et al. [12] introduced the BNLIT dataset and proposed a hybrid CNN-RNN model that combines Bidirectional RNN and LSTM models for image caption generation. Despite achieving good results using the BLEU (0.651) and METEOR (0.199) metrics, their dataset only contains one caption per image, which may limit the system's ability to generate coherent captions.

Humaira et al. [35] proposed a hybrid image captioning model using bidirectional long short-term memory (BiLSTM) and bidirectional gated recurrent unit (BiGRU). This study evaluated the hybrid method on two datasets (Flickr8k with 4000 and 8000 images), where the captions were translated into Bengali using Google Translate. They attained the most elevated BLEU-1 score of 0.661 and a METEOR score of 0.229. However, visual and textual attention could have been more focused. Ami et al. [15] introduced the attention mechanism for image captioning using the Flicker8K dataset. They utilized visual attention on the image, also known as spatial attention, and GRU as RNN to generate caption. The authors used pre-trained InceptionV3 and Xception to extract image features and used a soft attention mechanism to calculate the alignment score. This method showed the highest BLEU-1 score of 0.546 on Flicker8k-BN. The model's weaknesses were attributed to its dependence on culturally biased Flickr8k images and the production of captions that needed to be more fluent in Bengali. A recent study [8] proposed a transformer-based model for Bangla image captioning. They used InceptionV3 to extract image features and encoder-decoder architecture for caption generation. The authors employed multiple distinct datasets and attained a BLEU-1 score of 0.661 on BanglaLekha, 0.696 on Bornon, and 0.621 on merged Flickr8k-BN, Banglalekha, and Bornon datasets.

The current research differs from the existing studies in several ways. First, most past works used InceptionV3, VGG16, and Xception techniques for image feature extraction. Most past works employed pretrained CNN architectures such as InceptionV3, VGG16, and Xception for image feature extraction. However, all of the architectures are prone to vanishing gradient problems [36], making it difficult for a model to learn significant image-level features as weights of the earlier layers will not be updated. While other architecture is susceptible to the vanishing gradient problem, ResNet50 can eradicate this issue using the skip connection between the layers. Due to this advantage, we adopted ResNet50 architecture to extract the intricate image-level features in this work. Second, instead of the conventional encoder-decoder architecture like in past research, this work uses a context-aware attention mechanism to focus on the visual regions and generate meaningful descriptions selectively. Third, most past studies exclusively utilize the LSTM or GRU for decoding, which cannot capture information from future and past contexts [13,34,15,8]. To resolve this issue, this work adopted a bidirectional approach to effectively capture contextual dependencies from both directions. Lastly, no prior work in Bengali image captioning did not explore the adaptiveness of cross-domain transfer in Bengali image captioning.

3. Methodology

The primary objective of this work is to develop a model that can generate a caption for an image in Bengali. Given an image I and its associate descriptions, the proposed model generates a caption y as encoded as a sequence of 1-of-k encoded words $y = \{w_1, w_2, ...w_c\} \;\; w_i \; \epsilon \; R^k$, where k is the vocabulary size, and c is the caption length. Fig. 1 clearly depicts the stages involved in the research process. The first step involves importing the dataset, text tokenization, and image preprocessing. After that, embedded words are analyzed, and features are retrieved using ResNet-50. After applying context-aware attention and concatenating it with the preceding phases, BiGRU layer processing is performed, and captions are finally created. Fig. 2 depicts an overview of the proposed architecture of the Bengali image captioning framework. The proposed approach utilizes a pre-trained convolutional neural network (i.e., ResNet50) to extract the visual features as an encoder and a BiGRU model with the context-aware attention network to generate textual captions as a decoder.

3.1. Encoder: image features extraction

Before extracting the visual features from images, we preprocessed all the input images by resizing $224 \times 224 \times 3$ so that all images have the same size, and further feature extraction would be computationally less expensive. Keras preprocessing library has been used to preprocess the images. The image feature extraction portion has been illustrated in the upper block of Fig. 2. We have used a pre-trained convolutional neural network (ResNet-50) previously trained on ImageNet [37] dataset to extract the image features. We have used it to extract the global visual feature, $V = \{v_1, v_2, ...v_L\}$, which indicates the features extracted at different sub-regions of the image areas. Here, L (49) is the number of image regions, and v_i is the dimension of each region (which is 2048 in our case). The $7 \times 7 \times 2048$ vector size results from the downsampling process performed by the convolutional and pooling layers in the ResNet-50 architecture. The feature vector extracted from the image and the hidden state of the decoder are combined to generate the context vectors by the context-aware attention mechanism. However, before that, the dimension of the feature vector (v_i) is altered using 1×1 convolution to match the embedding dimension of 256 (described in Section 3.2.2) of the model. The embedding dimension is the same as the word embedding (Table 1).

3.2. Decoder: caption generation

The attention decoder is a crucial component of the image captioning model that generates textual descriptions of images. It incorporates visual information from different image regions into the caption-generation process. The decoder block consists of several constituent modules: Tokenization, embedding layer, context-aware attention, and GRU layer.

3.2.1. Text tokenization

In the first step, remove the special characters, punctuation symbols, and numbers from the captions of the texts. The numeric mapping of the words in the caption generates a preliminary vector representation of the captions. To get this mapping, a vocabulary (K) is created with k unique words (uw), $K = \{uw_1, uw_2, uw_3, ...uw_k\}$. The i^{th} word in a caption $y = [w_1, w_2, w_3, ...w_c]$ is replaced by the matching index number (i) of the words in vocabulary, K. By doing this, a caption is transformed into a sequence vector $s' = [w_{i1}, w_{i2}, w_{i3}, ...w_{ic}]$. However, the obtained sequence has a variable length (c), which is inappropriate for modeling. To resolve this issue, s' is transformed into a fixed length sequence vector $s = [w_{i1}, w_{i2}, w_{i3}, ...w_{il}]$ by padding where l is the maximum length

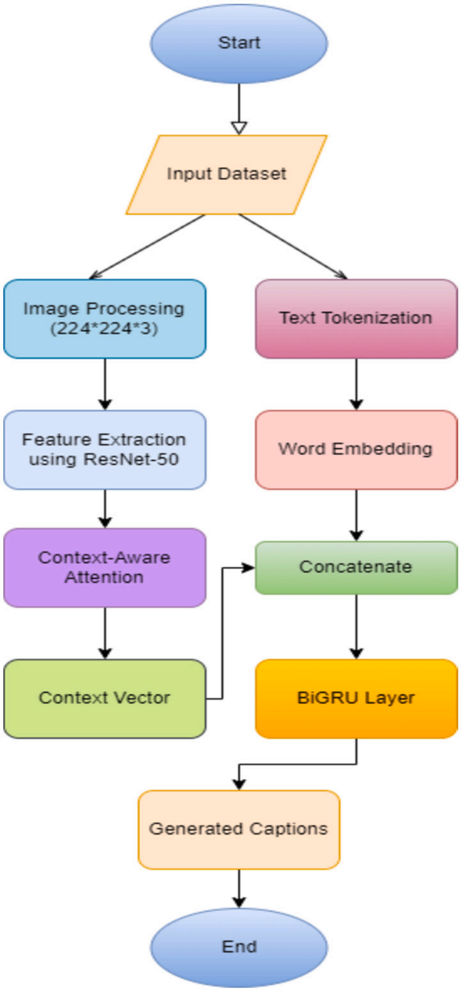


Fig. 1. Overview of the Image Captioning Model: From Image Preprocessing to Caption Generation.

of a caption. Padding involves adding special tokens (like $< PAD >$) to sequences shorter than the desired length. Fixed-length sequences aid in efficient memory management. Having sequences of equal length simplifies memory allocation and enables parallel data processing, which is essential for training models on extensive datasets.

3.2.2. Embedding layer

To encode the semantic information of the words in a caption to a global vector, each sequence vector in $S = [s_1, s_2, ..., s_N]$ passes through the Keras embedding layer to produce a word embedding vector (s_{em}) which is the embedding representation of the i^{th} sequence. We maintain the embedding dimension of size 256 to capture the relationship between words adequately.

3.2.3. Context aware attention

The attention model may, at any moment, determine the correct location in the image for image caption creation and produce a description that fits the observation content [38]. This work employed context-aware attention to capturing more precise sub-region image features instead of conventional RNN-based models [9,35], which only use the overall image features. The attention approach can adjust its gaze on corresponding areas of images when generating words related to various objects in an image, improving performance. This work uses The context-aware attention mechanism as the visual attention model.

The encoder provides the sub-region visual features of each image as output, $V = \{v_1, v_2, v_3, ..., v_L\}$; where $V_i \in \mathbb{R}^{2048}$. The previous decoder hidden state is the output of GRU denoted as h_{t-1}. The alignment between the sub-regions and words is calculated by Eq. (1).

$$e_{it} = W_{com}.\tanh(W_{encoder}.V_i + W_{decoder}.h_{t-1}) \tag{1}$$

Here, e_{it} is the aligned representation for the i-th image feature at time step t, and W is the trainable parameter.

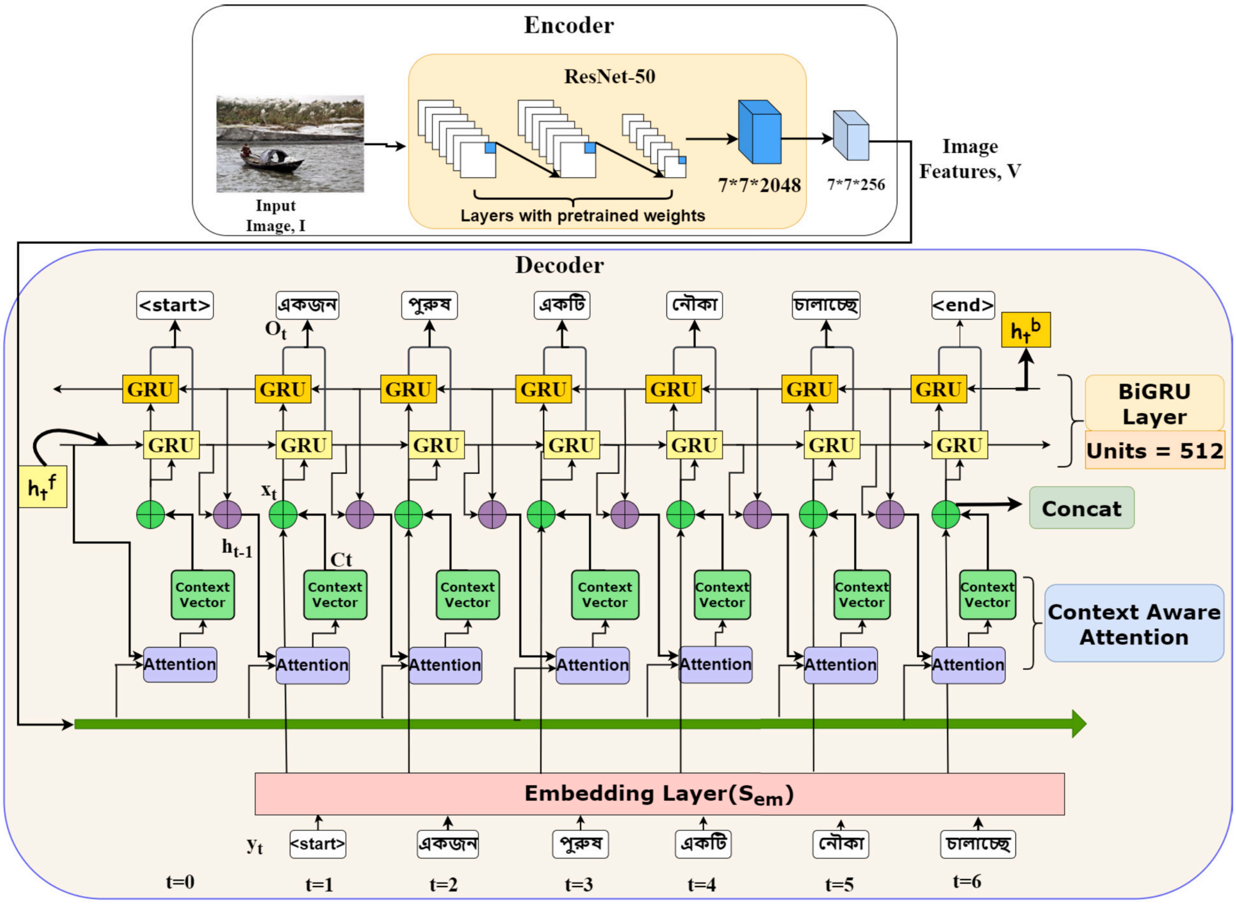


Fig. 2. An overview of the proposed encoder-decoder architecture for image captioning, where the upper block depicts image features (V) extraction by pre-trained ResNet-50 and the bottom block represents the decoder architecture incorporated with context-aware attention and bidirectional GRU.

The e_{it} will then be passed into a softmax function to calculate the attention weights α_{it} (Eqs. (2)-(3)).

$$\alpha_{it} = \frac{\exp(e_{it})}{\sum_{i=1}^{L} \exp(e_{it})} \tag{2}$$

$$c_t = \sum_{i=1}^{L} \alpha_{it}.V_i \tag{3}$$

where $\sum_{i=1}^{L} \alpha_{it} = 1$, the larger the weight of α_{it} is, the more significant the sub-region feature for generating a particular word. Finally, the context vector, c_t, is calculated as the weighted sum of V_i and α_{it}. This vector can offer precise visual guidance when generating words depending on a particular sub-region.

3.2.4. BiGRU layers

We employed a GRU network that utilizes two gates (reset and update). The main reason for choosing GRU as opposed to LSTM is that the GRU has fewer parameters and can be modeled more quickly. Let h_{t-1} be the previous hidden state of the decoder, $s_{em_{t-1}}$ be the embedding of the previous output word w_{t-1} derived from the embedding layer, and $c_t$is the attention context vector at time step t, computed as described in Eq. (4).

$$x_t = s_{em_{t-1}} \oplus c_t \tag{4}$$

Here, the input (x_t) to the BiGRU layer at time step t is the concatenation of the previous output word embedding ($s_{em_{t-1}}$) and the attention context vector (c_t).

The BiGRU layer computes the forward and backward hidden states $\mathbf{h}_t^f$ and $\mathbf{h}_t^b$ by processing the input x_t and previous decoder hidden state h_{t-1}. The update gate z_t, reset gate r_t, and candidate hidden state $\widetilde{h}_t$ are computed using Eqs. (5)-(12).

$$z_t^f = \sigma(W_z^f x_t + U_z^f h_{t-1}^f + U_z^b h_{t+1}^b + b_z^f) \tag{5}$$

$$r_t^f = \sigma(W_r^f x_t + U_r^f h_{t-1}^f + U_r^b h_{t+1}^b + b_r^f) \tag{6}$$

$$\tilde{h}_t^f = \tanh(W_h x_t + U_h(r_t^f \odot h_{t-1}^f) + b_h) \tag{7}$$

$$h_t^f = (1 - z_t^f) \odot h_{t-1}^f + z_t^f \odot \tilde{h}_t^f \tag{8}$$

$$z_t^b = \sigma(W_z^b x_t + U_z^b h_{t+1}^b + U_z^f h_{t-1}^f + b_z^b) \tag{9}$$

$$r_t^b = \sigma(W_r^b x_t + U_r^b h_{t+1}^b + U_r^f h_{t-1}^f + b_r^b) \tag{10}$$

$$\tilde{h}_t^b = \tanh(W_h x_t + U_h(r_t^b \odot h_{t+1}^b) + b_h) \tag{11}$$

$$h_t^b = (1 - z_t^b) \odot h_{t+1}^b + z_t^b \odot \tilde{h}_t^b \tag{12}$$

Where σ is the sigmoid function, $\odot$ is the element-wise multiplication for update and reset gate calculation, $W_z^f, W_r^f, W_h, W_z^b, W_r^b$ are weight matrices for the forward and backward GRU layers for input, x_t and $U_z^f, U_r^f, U_h, U_z^b, U_r^b$ are the weight matrices for the decoder hidden state, and $b_z^f, b_r^f, b_h, b_z^b, b_r^b$ are learnable bias vectors. The final decoder hidden state, h_t, is computed as a linear interpolation between the previous hidden state h_{t-1} and the candidate hidden state $\widetilde{h}_t$, controlled by the update gate z_t. The term $U_z^b h_{t+1}^b$ is the contribution of the backward hidden state, h_{t+1}^b to the update gate calculation at the time step, t. While it might seem counter-intuitive to use the backward hidden state in the forward pass calculation, it is important to note that the backward pass of the BiGRU layer operates in the forward pass's reverse direction. Therefore, the hidden state at a time step, $t+1$ in the backward pass, corresponds to the hidden state at a time step, $t-1$ in the forward pass. Thus, in Eq. (5) for z_t^f, the term $U_z^b h_{t+1}^b$ is effectively incorporated information from future time steps in the backward pass to inform the update gate calculation for the current time step in the forward pass.

The final hidden state at time step t is the concatenation of the forward and backward states. The process can be represented by Eq. (13).

$$h_t = h_t^f \oplus h_t^b \tag{13}$$

Here, h_t captures the information from both past and future words in the output sentence. At each time step, the BiGRU decoder layer concatenates the input word embedding with the attention context vector, which is a weighted sum of the image features based on the attention weights, to produce the hidden state h_t for the attention computation described in Eqs. (1)-(3).

3.3. Loss function

The sparse categorical cross-entropy loss function measures the difference between the predicted and actual word probability distributions. Given a batch of training examples, the loss function can be computed using Eq. (14).

$$L(y, \hat{y}) = -\frac{1}{N} \sum_{j=1}^{N} \sum_{l=1}^{L} y_{j,l} \log(\hat{y}_{j,l}) \tag{14}$$

Here, N is the batch size, L is the maximum sequence length, $y_{j,l}$ is the actual label for the l^{th} word in the j^{th} example, and $\hat{y}_{j,l}$ is the predicted probability distribution over the vocabulary for the l^{th} word in the j^{th} example.

To generate each word in the caption, the model takes as input the image features and the previous word in the sequence and predicts the probability distribution over the vocabulary for the next word. This process is repeated until the end-of-sequence token (<end>) is generated, indicating that the model has completed the caption. The caption generation is performed by Eq. (15).

$$P(y_t | y_{1:t-1}, V_i) = BiGRU(x_t \oplus h_{t-1}) \tag{15}$$

where $y_{1:t-1}$ are the previously generated words, V_i is the sub-region image feature vector, h_{t-1} is the previous decoder hidden state, and x_t is generated from context vector, c_t and previous word embedding vector described in Eq. (4).

The BiGRU computes the next hidden state and the probability distribution over the vocabulary using Eq. (16).

$$h_t, o_t = BiGRU(x_t \oplus h_{t-1}) \tag{16}$$

Here, h_t is the new hidden state, and o_t is the output of the decoder, which is a probability distribution over the vocabulary for the l^{th} word. During training, the loss function is used to compute the gradients of the model parameters concerning the loss used to update the parameters using ADAM. This process is repeated over multiple epochs until the model reaches a minimum loss.

Algorithm 1 presents the process of caption generation steps using the context-aware attention mechanism. Let us consider M is the total number of training samples, l is the maximum sequence length of captions, h is the number of hidden units in the BiGRU layer, v is the dimensionality of the image features, and e is the number of epochs. The computational complexity of this algorithm is computed as: (i) Image feature extraction requires $\mathcal{O}(M \times C)$ time where C is the computational complexity of extracting image features using ResNet-50, (ii) context-aware attention mechanism computes attention scores and context vectors which takes $\mathcal{O}(M \times l \times (h \times (l + v)))$ time complexity, (iii) BiGRU decoding involves the operations of matrix multiplications; therefore, this process takes $\mathcal{O}(M \times l \times h^2)$ complexity at each time step, (iv) finally, computing the loss function and updating model parameters take $\mathcal{O}(E \times M \times l)$. The overall complexity is dominated by the BiGRU decoding phase, and therefore, the time complexity of the context-aware attention algorithm is $\mathcal{O}(M \times l \times h^2)$.

Algorithm 1: Caption Generation using Context-Aware Attention Mechanism.

```
Input: Training set $\{(I_m, C_m)\}_{m=1}^{M}$, Vocabulary $K$ of size $k$, Pre-trained ResNet-50 model
Output: Generated captions $\{C_m\}_{m=1}^{M}$
Initialize image size $\leftarrow 224 \times 224 \times 3$ ;
Initialize embedding dim $\leftarrow 256$ ;
for each training sample $(I_m, C_m)$ do
    Extract image features, $V \leftarrow$ ResNet50 $(I_m)$ ;        // Feature vector $V = \{v_1, v_2, \ldots, v_{49}\}$
    Apply 1x1 convolution to $V$ to get dimension 256 ;
    Tokenize captions, map to indices from $K$ ;
    Pad sequences to fixed length $l$ ;
    $S \leftarrow$ Embedding layer $(l)$ ;        // Embedding dimension 256
    for each time step $t$ do
        $e_{it} \leftarrow W_{com} \cdot \tanh(W_{encoder} \cdot v_i + W_{decoder} \cdot h_{t-1})$ ;
        $\alpha_{it} \leftarrow \mathrm{softmax}(e_{it})$ ;
        $c_t \leftarrow \sum_{i=1}^{L} \alpha_{it} \cdot v_i$ ;        // Context vector
        $x_t \leftarrow \mathrm{concat}(sem_{t-1}, c_t)$ ;
        $h_t \leftarrow \mathrm{BiGRU}(x_t, h_{t-1})$ ;
    end
    for each word in the caption do
        $P(y_t | y_{1:t-1}, V) \leftarrow \mathrm{BiGRU}(x_t \oplus h_{t-1})$ ;
        $y_t \leftarrow \mathrm{argmax}(P)$ ;
    end
    Compute loss $L(y, \hat{y}) = -\frac{1}{N} \sum_{j=1}^{N} \sum_{l=1}^{L} y_{j,l} \log(\hat{y}_{j,l})$ ;
    Update parameters using ADAM optimizer ;
end
```

3.4. Training details

Keras Tuner is used to optimize hyperparameters of the proposed model, such as learning rate and batch size. First, a search space is configured with various values for each hyperparameter, including the optimizer and learning rate. The Hyperband search algorithm [39] determines the best values for the hyperparameters based on their impact on the validation set accuracy. This work did not consider other hyperparameters, such as the number of hidden units, number of BiGRU cells, and embedding dimension, as these were empirically selected. This approach can reduce computational costs. Table 1 demonstrates optimized hyperparameter values of the proposed model.

The proposed model utilizes the *Sparse Categorical Cross-Entropy* loss function and the Adam optimizer, with a learning rate of $10e^{-3}$. We used 64 instances per iteration during the training process and trained the model for 30 epochs. To prevent overfitting, we utilized the Keras checkpoint method, which continuously monitored the validation accuracy for five consecutive epochs and stopped the training process if there was no further improvement.

4. Experiments and analysis

We conduct experiments on the Google Colaboratory platform, which facilitates GPU usage. This work uses the NumPy (1.22.4) and pandas (1.3.5) libraries for data preparation and processing. TensorFlow and Keras (2.11.0) are used to implement each model. Scikit-learn (0.22.2) packages are used for model evaluation. The performance of the developed models is measured using the BLEU [40] and METEOR [41] scores considering a cross-domain transfer approach.

4.1. Datasets

Three publicly available Bengali datasets - (BAN-Cap [18], BanglaLekha [9] and Bornon [8]) - are utilized to assess the performance of the proposed method. The BAN-Cap [18] dataset comprises Bengali captions of the images in the Flickr8k dataset. The BAN-Cap dataset contained 8091 images, along with a total of 40,455 English-Bangla description pairs. The BanglaLekha [9] dataset featured a diverse range of 9154 images sourced from the public domain on the web. The images are primarily related to Bangladesh, with some relevance to the broader Indian Subcontinental context. However, the dataset poses a challenge, with only two captions available for each image, resulting in 18308 captions for the 9154 images. Consequently, the vocabulary size of this dataset is comparatively lower than the BAN-Cap dataset. The BAN-Cap dataset contains 12953 unique Bengali words, while the BanglaLekha dataset comprises only 6035. The Bornon [8] dataset is a collection of 4100 images, each with five corresponding captions with 20500 captions. The captions in the dataset predominantly contain Bengali words. Table 2 illustrates a brief overview of the datasets.

Table 1
Hyperparameters for training to the proposed model.

Hyperparameters	Optimum value
Embedding dimension	256
BiGRU hidden units	512
Optimizer	*Adam*
Learning rate	$10e^{-3}$
Batch size	64
Epochs	30
Loss function	*Sparse categorical crossentropy*

Table 2
Number of images and captions of each dataset.

Datasets	Total images	Total captions	Unique words
BAN-Cap [18]	8091	40455	12953
BanglaLekha [9]	9154	18308	6035
Bornon [8]	4100	20500	6228

4.2. Baselines

To compare the effectiveness of the proposed method, several encoder and decoder architectures are constructed to generate captions from images such as VGG16-attention-GRU, Inceptionv3-attention-GRU, ResNet50-attention, and ResNet50-hard attention-GRU. In the encoder, three state-of-the-art pre-trained CNN architectures, such as VGG16 [42], InceptionV3 [43] and ResNet-50 [44], are considered. Instead of relying on several hyper-parameters, VGG16 uses only 16 layers with weights. It is regarded as one of the top architectures for vision model systems. The inceptionV3 is an improved version of the fundamental model Inception V1, introduced in 2014 as GoogLeNet. The Inception V3 model optimized the network using many strategies for enhanced model adaption and computational efficiency. The ResNet-50 employs the bottleneck building block. A bottleneck residual block, often known as a “bottleneck”, reduces the number of parameters and matrix multiplications by using 1×1 convolutions. This significantly accelerates the training of each layer. Instead of using a stack of two levels, it employs three layers. Using pre-trained weights, images and features are extracted, omitting the last softmax layers for the classification. In the decoder, GRU has fewer parameters than LSTMs and, therefore, can be faster to train and require less memory. This is especially important in image captioning, where large amounts of data are processed. They have only two gates (reset and update) compared to three (input, output, and forget) in LSTM. Soft [24], and hard [45] both are considered to focus on the objects of images rather than whole image features.

5. Results

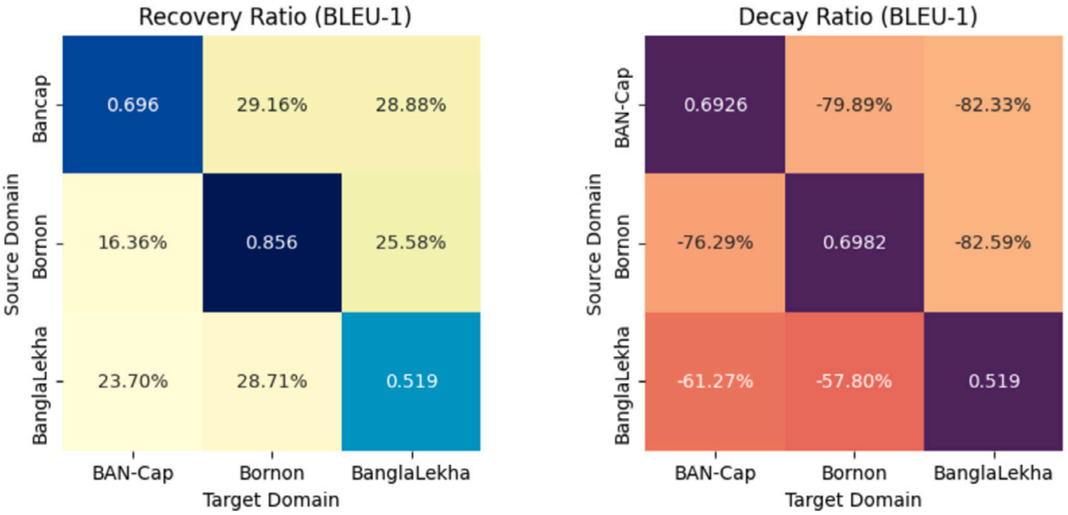
Table 3 compares the proposed and baseline methods on the test set of three datasets (BAN-Cap, BanglaLekha, and Bornon). In each dataset, five models are evaluated along with the proposed model. The first three models use VGG16, InceptionV3, and ResNet-50 as the image feature extractor, respectively, and attention mechanism with GRU for caption generation. Among these three models, the models having ResNet-50 achieved the best BLEU and METEOR score (0.662 for B-1 and 0.584 for M) while InceptionV3 performs slightly better ($\approx > 1\%$) than VGG-16 as encoder architecture. BLEU is valued for its simplicity and effectiveness in initial assessments but lacks the handling of synonyms. METEOR provides a detailed evaluation that considers recall, synonym matching, and word order, making it suitable for comprehensive assessments. Utilizing both BLEU and METEOR offers a well-rounded evaluation of captioning models, leveraging their strengths to assess different aspects of text quality. Meanwhile, the hard attention [45] with GRU as decoder architecture achieves the worst performance among these models (B-1 and M are less than 30% than ResNet50+Attention+GRU). The possible reason might be the incorporation of intricate (global) attention, which can not focus on the object-based features and thus fails to generate meaningful captions. However, the proposed method outperformed all the other baseline models, obtaining the highest BLEU score (B-1: 0.696) and METEOR score (0.584), which is 5% higher than the best baseline model (ResNet-50+Attention+GRU).

Similarly, for the BanglaLekha dataset, the proposed method again outperformed all other models across all metrics, achieving a BLEU-1, BLEU-4, and a METEOR score of 0.519, 0.322, and 0.471, respectively. The results indicate that the proposed method is better at generating more similar captions than ground truth captions. In the case of the Bornon dataset, intricate attention as decoder performs worst among all the baseline models but not as poor as the other two datasets because B-1 (0.784) is only 7% lower than the best baseline (ResNet+Attention+GRU) model performance (B-1: 0.846). However, the proposed model outdoes the best baseline model by ≈1% with 0.856 B-1 and 0.810 METEOR scores. This dataset’s evaluation scores are high because it uses five ground truth captions for short and crisp references.

Overall, the results demonstrate that the proposed approach, which utilizes context-aware attention with BiGRU, performs better than the baseline models across all three datasets. This superiority can be attributed to ResNet-50’s key feature of residual blocks, which allow for the efficient propagation of information through the architecture. Furthermore, using bidirectional GRU in the

Table 3
Performance evaluation of the proposed approach against different baseline models. Here, B denotes the BLEU score, and M represents METEOR. The best-performing system for each column is highlighted in boldface.

Dataset	Models	B-1	B-2	B-3	B-4	M
BAN-Cap	VGG16+Attention+GRU	0.619	0.524	0.458	0.372	0.501
	InceptionV3+Attention+GRU	0.627	0.516	0.433	0.373	0.502
	ResNet-50+Attention+GRU	0.662	0.569	0.457	0.394	0.577
	ResNet-50+Hard Attention+GRU	0.472	0.389	0.297	0.245	0.385
	Proposed Method	**0.696**	**0.612**	**0.523**	**0.456**	**0.584**
BanglaLekha	VGG16+Attention+GRU	0.428	0.333	0.265	0.224	0.375
	InceptionV3+Attention+GRU	0.445	0.346	0.277	0.236	0.391
	ResNet-50+Attention+GRU	0.501	0.418	0.349	0.306	0.466
	ResNet-50+Hard Attention+GRU	0.267	0.155	0.097	0.076	0.193
	Proposed Method	**0.519**	**0.434**	**0.369**	**0.322**	**0.471**
Bornon	VGG16+Attention+GRU	0.782	0.712	0.651	0.576	0.744
	InceptionV3+Attention+GRU	0.789	0.727	0.665	0.598	0.736
	ResNet-50+Attention+GRU	0.846	0.765	0.703	0.618	0.797
	ResNet-50+Hard Attention+GRU	0.784	0.685	0.612	0.502	0.691
	Proposed Method	**0.856**	**0.802**	**0.734**	**0.665**	**0.810**



(a) Recovery Ratio of BLEU-1

(b) Decay Ratio Of BLEU-1

Fig. 3. Cross-domain transfer analysis for Bengali image caption generation in terms of (a) recovery ratio and (b) decay ratio. The diagonal values in (a) indicate the BLEU-1 score when the source and target datasets are identical. In contrast, the off-diagonal values, representing cross-domain transfer performance, are given as percentages to indicate the relative change in performance due to domain shift.

decoder boosts the performance by processing the input captions in both directions, enabling the network to capture the context and dependencies between different words and generate more accurate captions.

5.1. Cross domain transfer

The ability of a model to generate captions for images in a domain different from its training domain is referred to as cross-domain transfer in image captioning. This work explores the cross-domain transfer by optimizing the model on a source domain (or dataset) and assessing on a target domain [46]. The recovery ratio [47] and decay ratio [47] are used to measure the performance of the cross-domain transfer. In the cross-domain experiments, we apply the proposed context-aware attention model with three datasets and only measure the ratio based on the BLEU-1 score. Fig. 3 shows the cross-domain transfer analysis for Bengali image caption generation in terms of (a) recovery ratio and (b) decay ratio.

The diagonal cells represent the BLEU-1 score when the source and target domains are identical. We point out that recovery and decay are separate indicators of domain transfer effectiveness. For instance, the domain transfer from BAN-Cap (English domain images) to Bornon (Bangla domain images) has a recovery rate of 16.36% and a decay ratio of -79.89%. The domain shifts from BAN-Cap to BanglaLekhaImageCaptions have a better recovery ratio (23.70%) and decay ratio (-57.80%). The maximum recovery ratio (29.16%) and lowest decay ratio (-61.27%) are achieved when the domain shifts from BanglaLekhaImageCaptions to the BAN-

Cap dataset. However, domain transfer performance is notably low due to the datasets' stark differences. The BAN-Cap dataset is an extension of the Flickr8k dataset, consisting of 8000 photos and 40000 English-Bangla caption pairs. Native speakers write captions in English and Bangla, yet the dataset lacks cultural diversity due to its Western inclination in the English domain. In contrast, the BanglaLekha-ImageCaptions dataset is substantially smaller, with only 9,154 images acquired from the public domain, covering a broad range of subjects primarily related to Bangladesh and the larger Indian Subcontinental environment.

Although two Bangla domain datasets, Bornon and BanglaLekha-ImageCaptions, should have a higher recovery ratio and lower decay ratio, performance is almost similar to BAN-cap. The reason behind this can be attributed to the fact that the image sources in BanglaLekhaImageCaptions are vastly different from Bornon. Moreover, BanglaLekha-ImageCaptions has more extended captions focusing on the broader Indian Subcontinental culture, while Bornon has more concise captions based on personal photography club. For these reasons, cross-domain transfer for image captioning is more difficult as different datasets have different culture-rooted images in their domain.

5.2. Qualitative analysis

We compare the proposed method to the best two baseline methods to better understand how different approaches work. Table 4 and Table 5 demonstrate a detailed comparison. The proposed model outperformed the human judgments concerning the evaluation scores and caption quality.

In Table 4, it is observed that the proposed model with the context-aware attention network predicted the animal 'sheep' in sample 1 accurately, while the other two baseline methods failed to do so. As expected, the proposed model achieved significantly better evaluation scores (BLEU-1: 0.80 and METEOR: 0.768). For sample 2, although the model's evaluation scores were lower than the other two baseline methods, it attempted to predict the detailing of objects in the test image rather than providing concise predictions. It tried to predict the jersey color of hockey players but could not produce the exact sequence of words, and a BLEU-1 score of 0.66 is quite good. The predicted captions and evaluation scores were almost identical for all three methods in sample 3 (shown in Table 5), where none of the methods could predict the exact number of men visible in the image due to their small size. In the case of sample 4 (shown in Table 5), the proposed model outperformed both baseline methods regarding evaluation scores and almost matched the reference caption with BLEU-1:0.80. The results and analysis revealed that the proposed model can generate precise and accurate image captions useful in various applications. Overall, the findings suggest that the proposed model is a promising image captioning approach and could outperform existing methods.

To facilitate a better understanding of the suggested model, the process of image captioning involved the display of visualization for context-aware attention weights. Table 6 illustrates the plotting of attention weights that help determine which image area was targeted during the caption generation. It also represents the significant feature of each image in highlighted form with the predicted word of the relevant vital features. This representation demonstrates that not every aspect of the image is required to anticipate a caption. Furthermore, if the model does not apply this type of attention architecture, it may lose several object details.

As shown in Table 6, the proposed model demonstrates the ability to predict precise words by focusing on the significant features of the image object rather than emphasizing the entire aspect. This approach enhances the accuracy of the generated caption by ensuring that the relevant features are adequately highlighted. However, in sample 1, the proposed method struggles to highlight the 'Bangladeshi flag' and 'cloud' objects effectively, resulting in lower evaluation scores (BLEU-1: 0.25). In sample 2, the proposed model perfectly focuses on the objects, resulting in a perfect evaluation score (All BLEU: 1.00). This observation underscores the importance of efficiently detecting noteworthy features and objects to generate high-quality captions.

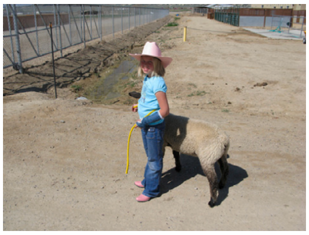
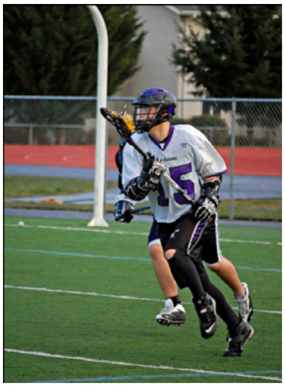
5.3. Comparison with existing models

In this study, we compared the performance of our proposed method with existing models, such as the CNN-CNN merge architecture, visual attention, encoder-decoder architecture, and Transformer, using machine-translated evaluation metrics, including BLEU and METEOR, on three different datasets. While BanglaLekha ImageCaptions and BAN-Cap are widely used, Bornon is a relatively new dataset, and fewer notable experiments have yet to be performed. The results, shown in Table 7, demonstrate that our context-aware attention-based model outperforms all existing methods for BAN-Cap and Bornon dataset in the Bangla language domain, surpassing the visual attention [15] and transformer [8] architecture. In particular, the proposed method achieves a BLEU-1 score of 0.696 on the BAN-Cap dataset and 0.856 on the Bornon dataset, significantly higher than the other existing models' scores. Similarly, when looking at the METEOR scores, the proposed method again outperforms the other models on all three datasets. In particular, the proposed method achieves a METEOR score of 0.584 on the BAN-Cap dataset, surpassing 0.297 by Transformer, 0.471 on the BanglaLekha dataset, and 0.810 on the Bornon dataset, which is significantly higher than the visual-attention model and Transformer. In the visual attention model [15], they executed InceptionV3 as encoder architecture and soft attention with GRU as decoder architecture. Shah et al. [8] obtained a transformer with three layers for the Bornon dataset to get the best evaluation scores. CNN-CNN merge architecture by Fayaz et al. [34] used a multimodal approach with ResNet-50 as encoder and 1D-CNN as decoder architecture.

5.4. Discussion

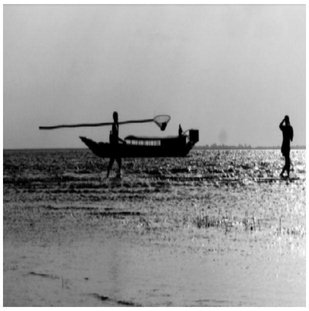
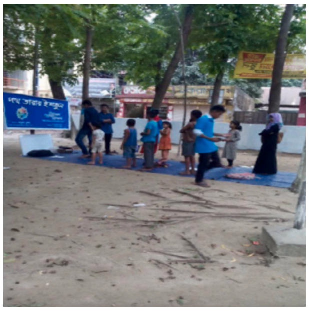
Although the proposed method had lower BLEU scores compared to the CNN-CNN Merge and transformer models for the BanglaLekhaImageCaptions dataset, the quality of the generated captions was still good, as demonstrated in Table 5. BLEU score relies solely on word overlap between the generated and reference captions, and it does not consider the fluency and coherence of

Table 4
Comparison of the performance and the quality of the captions generated by the proposed method and the two best baseline methods for the Ban-Cap dataset. Here, the red color indicates poor performance, while the blue color indicates a superior outcome.

Sample Image	Predicted Caption with Evaluation Score
 (1) Test Image from BAN-Cap	**Reference Caption:** কাউবয় টুপি পড়া মেয়েটি একটি ভেড়া নিয়ে দাড়িয়ে আছে, একটি গোলাপি টুপি পড়া এক মেয়ে ধূলিকণা পথে দাড়িয়ে আছে **Baseline Method-1: InceptionV3 + Attention + GRU** **Predicted Caption:** এক মহিলা একটি বাদামি রংয়ের জিনস এবং তার একটি বাদামি রংয়ের একটি জিনিস ধূলিকণা নিয়ে একটি বাদামি রংয়ের একটি জিনিস ধূলিকণা নিয়ে... BLEU-1: 0.046728 BLEU-4: 0.004763 Meteor Score: 0.072816 **Baseline Method-2: Resnet50 + Attention + GRU** **Predicted Caption:** একটি মেয়ে ধূলিকণা পথে ধরে আছে (A girl is holding dust on the road) BLEU-1: 0.716531 BLEU-4: 0.432982 Meteor Score: 0.377976 **Proposed Method: Resnet50 + Attention + BiGRU** **Predicted Caption:** গোলাপি রংয়ের পোশাক পরা মেয়েটি একটি ভেড়া নিয়ে দাড়িয়ে আছে (A girl in a pink dress is standing with a sheep) BLEU-1: 0.800000 BLEU-4: 0.660632 Meteor Score: 0.768109
 (2) Test Image from BAN-Cap	**Reference Caption:** একটি হকি খেলা হচ্ছে, বেগুনি সাদা এবং কালো রঙের একটি খেলোয়াড় মাঠের শেষের কাছে একটি নাটক তৈরি করার চেষ্টা করে **Baseline Method-1: InceptionV3 + Attention + GRU** **Predicted Caption:** দুজন হকি খেলা হচ্ছে (Two are playing hockey) BLEU-1: 1.000000 BLEU-4: 0.181252 Meteor Score: 0.755932 **Baseline Method-2: ResNet50 + Attention + GRU** **Predicted Caption:** খেলার পোশাকে দুজন খেলোয়াড় মাঠে দৌড়াচ্ছে (Two players in sportswear are running on the field) BLEU-1: 1.000000 BLEU-4: 0.181252 Meteor Score: 0.755932 **Proposed Method: ResNet50 + Attention + BiGRU** **Predicted Caption:** বেগুনি সাদা কালো রংের খেলা খেলার সময় একটি খেলার বল মাঠে ছুড়ে মারলো (Throwing a game ball on the field during a game of purple white black color) BLEU-1: 0.666666 BLEU-4: 0.803014 Meteor Score: 0.285897

the sentences. Moreover, the BanglaLekhaImageCaptions dataset has only two reference captions, unlike the BAN-Cap and Bornon datasets, which have five reference captions. As a result, the number of unique words is lower, ultimately leading to lower BLEU scores. However, the proposed method still outperformed the other models regarding the qualitative analysis (presented in Section 5.2), which

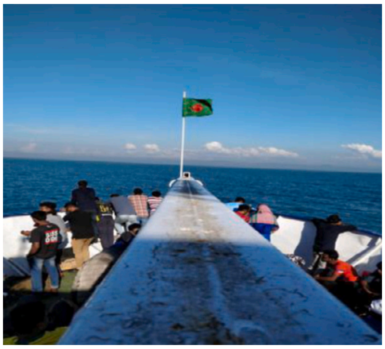
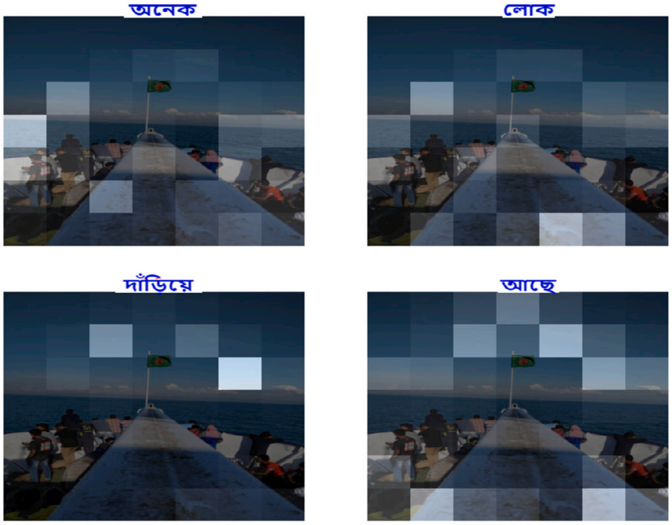
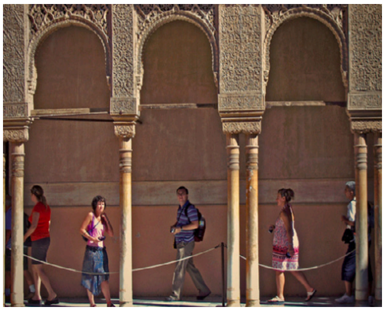
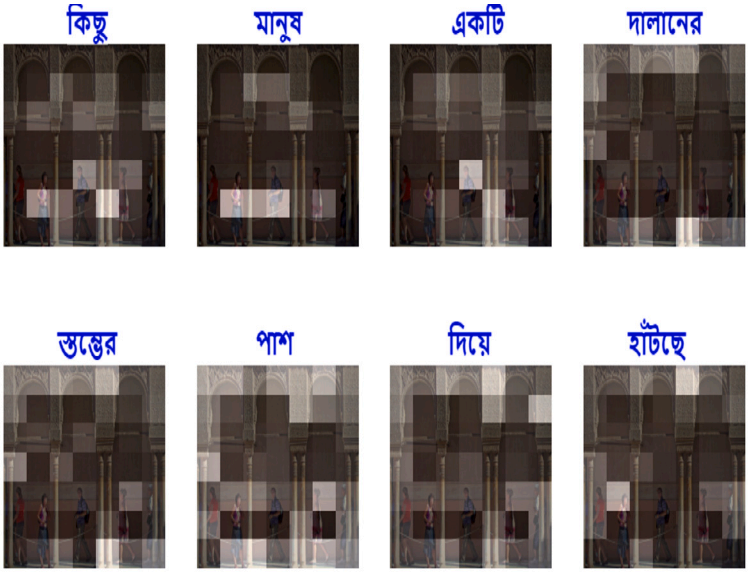
Table 5
Comparison of the performance and quality of the captions generated by the proposed method with the best two baseline methods for the Banglalekha dataset. Here, the red color indicates poor performance, while the blue color indicates a superior outcome.

Sample Image	Predicted Caption with Evaluation Score
 (3) Test Image from Banglalekha	**Reference Caption:** একটি নৌকায় মানুষ আছে পানিতে দুইজন মানুষ আছে, নদীর পানিতে চলে যাচ্ছে একটি নৌকা এবং ২ জন পুরুষ পানিতে হেঁটে যাচ্ছে **Baseline Method-1: InceptionV3 + Attention + GRU** **Predicted Caption:** তিনটি নৌকায় কয়েকজন মানুষ আছে (There are some people in three boats) BLEU-1: 0.329286 BLEU-4: 0.073618 Meteor Score: 0.331890 **Baseline Method-2: ResNet50 + Attention + GRU** **Predicted Caption:** একটি নৌকায় কয়েকজন মানুষ দেখা যাচ্ছে (A few people are seen in a boat) BLEU-1: 0.477687 BLEU-4: 0.082805 Meteor Score: 0.327635 **Proposed Method: ResNet50 + Attention + BiGRU** **Predicted Caption:** একটি নৌকার উপর কয়েকজন মানুষ আছে (There are some people on a boat) BLEU-1: 0.358265 BLEU-4: 0.077058 Meteor Score: 0.327635
 (4) Test Image from Banglalekha	**Reference Caption:** কয়েকজন মানুষ দাড়িয়ে আছে, কয়েকজন ছোটো ছেলেমেয়েকে পড়াশুনা করাচ্ছে **Baseline Method-1: InceptionV3 + Attention + GRU** **Predicted Caption:** কয়েকজন মানুষ হেঁটে যাচ্ছে (Some people are walking) BLEU-1: 0.500000 BLEU-4: 0.168218 Meteor Score: 0.468750 **Baseline Method-2: ResNet50 + Attention + GRU** **Predicted Caption:** কয়েকজন মানুষ হেঁটে যাচ্ছে (Some people are walking) BLEU-1: 0.500000 BLEU-4: 0.168218 Meteor Score: 0.468750 **Proposed Method: ResNet50 + Attention + BiGRU** **Predicted Caption:** সামনে কয়েকজন মানুষ দাড়েয়ে আছে (A few people are standing in front) BLEU-1: 0.800000 BLEU-4: 0.668740 Meteor Score: 0.967988

indicates its effectiveness in generating meaningful and coherent captions for the BanglaLekhaImageCaptions dataset. These findings suggest the proposed model is a promising approach for generating high-quality image captions in Bengali.

For future prospects, combining reinforcement learning (RL) [48,49] and fuzzy systems [50] can leverage the strengths of image caption generation. Several works [51–56] employ this approach in various research areas. RL will allow for optimizing the caption generation process through trial and error, while fuzzy systems can handle uncertainty and incorporate human-like reasoning into the

Table 6
Visualization of attention weights for a clearer understanding of how the model predicts each word by focusing on objects from test images.

Sample Image and Evaluation score	Visualization of attention weights
 **Reference Caption:** আকাশে মেঘ আছে, একটি পতাকা আছে **Predicted Caption:** অনেক লোক দাড়িয়ে আছে (Many people are standing) BLEU-1: 0.250000 BLEU-2: 0.107482 BLEU-3: 0.066131 BLEU-4: 0.061033 Meteor Score: 0.161290	
 **Reference Caption:** একদল লোক একটি প্রদর্শনীতে হাটঁছে, পর্যটকরা ঐতিহাসিক স্থান ভ্রমণ করছেন **Predicted Caption:** কিছু মানুষ একটি দালানের স্তম্ভের পাশ দিয়ে হাঁটছে (Some people are walking by the pillars of a building) BLEU-1: 1.000000 BLEU-2: 1.000000 BLEU-3: 1.000000 BLEU-4: 1.000000 Meteor Score: 0.99023	 কিছু মানুষ একটি দালানের স্তম্ভের পাশ দিয়ে হাঁটছে

model. This combination can lead to more accurate and contextually appropriate captions. However, this combination will require significant computational complexity compared to our proposed context-aware attention method.

Table 7
Comparison of the performance of the proposed model with existing models. Here, B represents the BLEU score, and M represents METEOR. The best-performing system for each column is highlighted in boldface. The optimal performance system of the BanglaLekha dataset BLEU-1, 2, 3, 4 are highlighted in boldface.

Dataset	Models	B-1	B-2	B-3	B-4	M
BAN-Cap	CNN-CNN Merge [18]	0.565	0.355	0.221	0.131	0.281
	Visual-Attention [18]	0.587	0.368	0.251	0.144	0.293
	Transformer [18]	0.623	0.396	0.251	0.152	0.297
	proposed method	**0.696**	**0.612**	**0.523**	**0.456**	**0.584**
BanglaLekha	CNN-CNN Merge [34]	0.651	0.426	0.272	0.175	0.297
	Visual-Attention [15]	0.57	0.46	0.39	0.32	0.21
	Transformer [8]	**0.665**	**0.556**	**0.476**	**0.408**	0.255
	proposed method	0.519	0.434	0.369	0.322	**0.471**
Bornon	Visual-Attention [15]	0.605	0.492	0.408	0.351	0.348
	Transformer [8]	0.696	0.589	0.507	0.439	0.361
	proposed method	**0.856**	**0.802**	**0.734**	**0.665**	**0.810**

6. Conclusion

This study presents a context-aware, attention-based image captioning system for Bengali. It leverages pre-trained ResNet-50 to extract subregion-based visual features. Combining these features with context-aware attention and a BiGRU layer, the model constructs high-performance descriptions in Bangla. Evaluation with BLEU and METEOR metrics demonstrates superior performance over baseline and current models on two image caption datasets, except BanglaLekhaImageCaption, where the limited reference captions impact evaluation scores. Despite this, comparative analysis showcases the proposed approach's significant outperformance of current state-of-the-art models. However, there is room for improvement, particularly in addressing potential dataset bias and integrating semantic attention. Future work aims to incorporate dual attention, encompassing visual and semantic attention, to enhance descriptive accuracy and explore language adaptability within the model architecture.

CRediT authorship contribution statement

Ahatesham Bhuiyan: Writing – original draft, Software, Methodology, Data curation, Conceptualization. **Eftekhar Hossain:** Writing – original draft, Supervision, Resources, Formal analysis, Conceptualization. **Mohammed Moshiul Hoque:** Writing – review & editing, Validation, Project administration, Investigation. **M. Ali Akber Dewan:** Writing – review & editing, Validation, Software, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) under the grant number RGPIN-2020-06080 and the Directorate of Research and Extension (DRE), Chittagong University of Engineering & Technology (CUET), Bangladesh (CUET/DRE/2023-24/CSE/025).

References

[1] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D.A. Shamma, et al., Visual genome: connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vis. 123 (2017) 32–73.
[2] S. Khalid, S. Wu, A. Wahid, A. Alam, I. Ullah, An effective scholarly search by combining inverted indices and structured search with citation networks analysis, IEEE Access 9 (2021) 120210–120226.
[3] J. Kim, A. Rohrbach, T. Darrell, J. Canny, Z. Akata, Textual explanations for self-driving vehicles, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 563–578.
[4] A. Goyal, M. Mandal, V. Hassija, M. Aloqaily, V. Chamola, Captionomaly: a deep learning toolbox for anomaly captioning in social surveillance systems, IEEE Trans. Comput. Soc. Syst. (2023).
[5] M. Hodosh, P. Young, J. Hockenmaier, Framing image description as a ranking task: data, models and evaluation metrics, J. Artif. Intell. Res. 47 (2013) 853–899.
[6] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, C.L. Zitnick, Microsoft coco captions: data collection and evaluation server, arXiv preprint, arXiv:1504.00325, 2015.

[7] Y. Hirota, Y. Nakashima, N. Garcia, Quantifying societal bias amplification in image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13450–13459.
[8] F.M. Shah, M. Humaira, M.A.R.K. Jim, A.S. Ami, S. Paul, Bornon: Bengali image captioning with transformer-based deep learning approach, arXiv preprint, arXiv:2109.05218, 2021.
[9] M. Rahman, N. Mohammed, N. Mansoor, S. Momen, Chittron: an automatic bangla image captioning system, Proc. Comput. Sci. 154 (2019) 636–642.
[10] S. Herdade, A. Kappeler, K. Boakye, J. Soares, Image captioning: transforming objects into words, Adv. Neural Inf. Process. Syst. 32 (2019).
[11] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: a neural image caption generator, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.
[12] M.A. Jishan, K.R. Mahmud, A.K. Al Azad, M.S. Alam, A.M. Khan, Hybrid deep neural network for bangla automated image descriptor, J. Adv. Comput. Intell. Inform. 6 (2) (2020) 109–122.
[13] A.H. Kamal, M.A. Jishan, N. Mansoor, Textmage: the automated bangla caption generator based on deep learning, in: 2020 International Conference on Decision Aid Sciences and Application (DASA), IEEE, 2020, pp. 822–826.
[14] T. Deb, M.Z.A. Ali, S. Bhowmik, A. Firoze, S.S. Ahmed, M.A. Tahmeed, N. Rahman, R.M. Rahman, Oboyob: a sequential-semantic Bengali image captioning engine, J. Intell. Fuzzy Syst. 37 (6) (2019) 7427–7439.
[15] A.S. Ami, M. Humaira, M.A.R.K. Jim, S. Paul, F.M. Shah, Bengali image captioning with visual attention, in: 2020 23rd International Conference on Computer and Information Technology (ICCIT), IEEE, 2020, pp. 1–5.
[16] M.A.H. Palash, M.A.A. Nasim, S. Saha, F. Afrin, R. Mallik, S. Samiappan, Bangla image caption generation through cnn-transformer based encoder-decoder network, in: Proceedings of International Conference on Fourth Industrial Revolution and Beyond 2021, Springer, 2022, pp. 631–644.
[17] D. Vasić, M.K. Vasić, Syntax-aware neural semantic role labeling for morphologically rich languages, in: 2020 International Conference on Software, Telecommunications and Computer Networks (SoftCOM), IEEE, 2020, pp. 1–6.
[18] M.F. Khan, S. Shifath, M.S. Islam, Ban-cap: a multi-purpose English-bangla image descriptions dataset, arXiv preprint, arXiv:2205.14462, 2022.
[19] B.A. Plummer, L. Wang, C.M. Cervantes, J.C. Caicedo, J. Hockenmaier, S. Lazebnik, Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2641–2649.
[20] R. Socher, A. Karpathy, Q.V. Le, C.D. Manning, A.Y. Ng, Grounded compositional semantics for finding and describing images with sentences, Trans. Assoc. Comput. Linguist. 2 (2014) 207–218.
[21] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–2634.
[22] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3128–3137.
[23] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, arXiv preprint, arXiv:1406.1078, 2014.
[24] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint, arXiv:1409.0473, 2014.
[25] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention, in: International Conference on Machine Learning, PMLR, 2015, pp. 2048–2057.
[26] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image captioning with semantic attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4651–4659.
[27] Y. Li, Y. Pan, T. Yao, T. Mei, Comprehending and ordering semantics for image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17990–17999.
[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.
[29] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: consensus-based image description evaluation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4566–4575.
[30] Z. Fang, J. Wang, X. Hu, L. Liang, Z. Gan, L. Wang, Y. Yang, Z. Liu, Injecting semantic concepts into end-to-end image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18009–18019.
[31] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, P. Anderson, Nocaps: novel object captioning at scale, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8948–8957.
[32] P. Sharma, N. Ding, S. Goodman, R. Soricut, Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2556–2565.
[33] J. Luo, Y. Li, Y. Pan, T. Yao, J. Feng, H. Chao, T. Mei, Semantic-conditional diffusion networks for image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23359–23368.
[34] M. Faiyaz Khan, S. Sadiq-Ur-Rahman, M. Saiful Islam, Improved Bengali image captioning via deep convolutional neural network based encoder-decoder model, in: Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020, Springer, 2021, pp. 217–229.
[35] M. Humaira, P. Shimul, M.A.R.K. Jim, A.S. Ami, F.M. Shah, A hybridized deep learning method for Bengali image captioning, Int. J. Adv. Comput. Sci. Appl. 12 (2) (2021).
[36] S. Hochreiter, The vanishing gradient problem during learning recurrent neural nets and problem solutions, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 6 (02) (1998) 107–116.
[37] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.
[38] N.D. Bruce, J.K. Tsotsos, Saliency, attention, and visual search: an information theoretic approach, J. Vis. 9 (3) (2009) 5.
[39] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, A. Talwalkar, Hyperband: a novel bandit-based approach to hyperparameter optimization, J. Mach. Learn. Res. 18 (1) (2017) 6765–6816.
[40] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[41] S. Banerjee, A. Lavie, Meteor: an automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005, pp. 65–72.
[42] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint, arXiv:1409.1556, 2014.
[43] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[44] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[45] M.-T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, arXiv preprint, arXiv:1508.04025, 2015.
[46] I. Turc, K. Lee, J. Eisenstein, M.-W. Chang, K. Toutanova, Revisiting the primacy of English in zero-shot cross-lingual transfer, arXiv preprint, arXiv:2106.16171, 2021.
[47] C. Toraman, F. Şahinuç, E.H. Yilmaz, Large-scale hate speech detection with cross-domain transfer, arXiv preprint, arXiv:2203.01111, 2022.
[48] D. Zhang, W. Wang, J. Zhang, T. Zhang, J. Du, C. Yang, Novel edge caching approach based on multi-agent deep reinforcement learning for Internet of vehicles, IEEE Trans. Intell. Transp. Syst. (2023).

[49] J. Zhang, M.-j. Piao, D.-g. Zhang, T. Zhang, W.-m. Dong, An approach of multi-objective computing task offloading scheduling based nsgs for iov in 5g, Clust. Comput. 25 (6) (2022) 4203–4219.

[50] D.-G. Zhang, C.-H. Ni, J. Zhang, T. Zhang, Z.-H. Zhang, New method of vehicle cooperative communication based on fuzzy logic and signaling game strategy, Future Gener. Comput. Syst. 142 (2023) 131–149.

[51] D.-g. Zhang, X. Wang, X.-d. Song, New medical image fusion approach with coding based on scd in wireless sensor network, J. Electr. Eng. Technol. 10 (6) (2015) 2384–2392.

[52] T. Zhang, D.-g. Zhang, H.-r. Yan, J.-n. Qiu, J.-x. Gao, A new method of data missing estimation with fnn-based tensor heterogeneous ensemble learning for Internet of vehicle, Neurocomputing 420 (2021) 98–110.

[53] D.-G. Zhang, W.-M. Dong, T. Zhang, J. Zhang, P. Zhang, G.-X. Sun, Y.-H. Cao, New computing tasks offloading method for mec based on prospect theory framework, IEEE Trans. Comput. Soc. Syst. (2022).

[54] D.-g. Zhang, J. Zhang, C.-h. Ni, T. Zhang, P.-z. Zhao, W.-m. Dong, New method of edge computing based data adaptive return in Internet of vehicles, IEEE Trans. Ind. Inform. (2023).

[55] D. Zhang, G. Li, K. Zheng, X. Ming, Z.-H. Pan, An energy-balanced routing method based on forward-aware factor for wireless sensor networks, IEEE Trans. Ind. Inform. 10 (1) (2013) 766–773.

[56] D. Zhang, X. Wang, X. Song, D. Zhao, A novel approach to mapped correlation of id for rfid anti-collision, IEEE Trans. Serv. Comput. 7 (4) (2014) 741–748.