ORIGINAL RESEARCH

A Visual Attention-Based Model for Bengali Image Captioning

Bidyut Das[1] · Ratnabali Pal[2,3] · Mukta Majumder[4] · Santanu Phadikar[5] · Arif Ahmed Sekh[6]

Received: 27 June 2022 / Accepted: 3 January 2023

Abstract

Image caption or description generation is a fundamental task that involves computer vision (CV) and natural language processing (NLP) ideas to recognize an image-context and produces description(s) using a natural language. Bengali is one of the world's most commonly spoken languages, ranking fifth. For this reason, large research achievements have been recognized to image captioning, i.e. explaining images with grammatically correct and semantically meaningful Bengali sentences. Many established datasets exist for image caption generation in English, but no standard dataset is available for Bengali. This paper proposes a model for generating automatic image captions in the Bengali language. This study uses only two initial available Bengali datasets to train the encoder-decoder neural network model. We have curated the human errors available in the datasets and benchmarked. Experimental results of this proposed model performs better than other baseline models using these datasets. It achieved 0.67 and 0.65 BLEU-1, and 0.26 and 0.24 BLEU-4 respectively. We expect that our research will take attentions to more researchers from regional language understanding and accelerate Bengali vision-language understanding.

Keywords Bengali image caption · Image description · Bengali NLP · Deep learning · Attention mechanism

✉ Bidyut Das
bidyut2002in@gmail.com

Ratnabali Pal
pal.ratnabali@gmail.com

Mukta Majumder
mukta_jgec_it_4@yahoo.co.in

Santanu Phadikar
sphadikar@yahoo.com

Arif Ahmed Sekh
skarifahmed@gmail.com

[1] Department of Information Technology, Haldia Institute of Technology, Haldia, West Bengal 721657, India

[2] Department of Applied Mathematics, NIT Durgapur, Durgapur, West Bengal 713213, India

[3] Department of Computational Science, Brainware University, Kolkata, West Bengal 700125, India

[4] Department of Computer Science and Technology, University of North Bengal, Siliguri, West Bengal 734014, India

[5] Department of Computer Science and Engineering, Maulana Abul Kalam Azad University of Technology, W.B., Haringhata, West Bengal 741249, India

[6] School of Computer Science and Engineering, XIM University, Bhubaneswar, Odisha 752050, India

Introduction

Generating image description from an image is a demanding task. This task is complicated compared to image recognition and image classification [1]. The description of an image depends on not only the image objects but also their relationships, activities, and attributes. Hence, the language model is required to express the semantic knowledge of the natural language. Image description can be beneficial to visually impaired people for interpreting massive content on the web network. It also helps people to recognize and navigate through a wide range of unstructured visual information. The advanced research on object identification and language modeling makes it possible to produce relevant image captions. Automated textual description generating is now frontier research area in field of image and natural-language processing [2].

The automatic image caption generation has received a lot of interest in recent years. Several researchers worked on image description generation and found that employing the encoder–decoder framework improved the result of description generation significantly [3]. The encoder–decoder framework employs both convolutional and recurrent neural networks (CNN + RNN). The image

Published online: 16 February 2023

is encoded using CNN as an encoder while decoded using RNN as a decoder to predict the word sequences [4–6] (Fig. 1).

Several datasets for English captioning are freely available in the literature, such as Flickr8K [7], Flickr30K [8], MSCOCO [9] and many more. However, due to a lack of regional language datasets, we have only located a few works of image captioning in regional languages. Image searching, image descriptions for social media, news, and e-commerce websites, and audio from image descriptions for visually impaired persons are all viable uses of image captioning in the regional language. English, Chinese, and Spanish are the three most widely used languages for image captioning research [10]. However, there has been few notable research algorithm based on AI techniques on image caption generating in the Bengali language. It has been recognized to be world's fifth most common language while few native Bengali speakers here don't understand English well. As a result, Bengali image captioning can be helpful to those people who are unable to speak and understand other languages besides Bengali. The article's key contributions are as follows:

- We have curated and benchmarked the available Bengali datasets for image caption generation. The available datasets are erroneous and not standardized or benchmarked. We have verified the datasets and corrected by human annotators.
- We have benchmarked the datasets varying by state-of-the-art CNN-based encoders and RNN-based decoders. The study helps us to find a suitable method for the Bengali image captioning system.
- We have developed an image description model in Bengali using the attention-based encoder–decoder. We have used CNN-based encoder for feature extraction and GRU-based decoder for description generation.
- We introduced a novel loss function combining the CNN-based attention loss and GRU-based decoder loss for the task.

There are six sections in this paper. The past studies on English image captioning and captioning on other regional languages are described in Section "Related Work". Section "Proposed Method" explains our proposed method. Following that, section "Dataset" shows the utilized datasets for this experiment and the training approach to generate Bengali image captions. The experiment's results and discussion are shown in section "Experimental Results with Discussion". Finally, Section "Conclusion" concludes this work.

Related Work

This section presents various strategies and state-of-the-art methods applied in the image captioning literature. Numerous researchers worked on this topic and provided many approaches for creating image captions [11]. We describe here the earlier works into two subsections depending on the image-captioning language: image captioning in English (Section "Image Caption Generation in English") and captioning in other regional languages (2.2). We summarize the newest articles and latest achievements from the previous popular methodologies in image captioning literature at the end of this Sect. Our Observation from the Earlier Literature.

Fig. 1 An attention provides the clue to construct the sentence, and RNN-based decoder constructs the caption

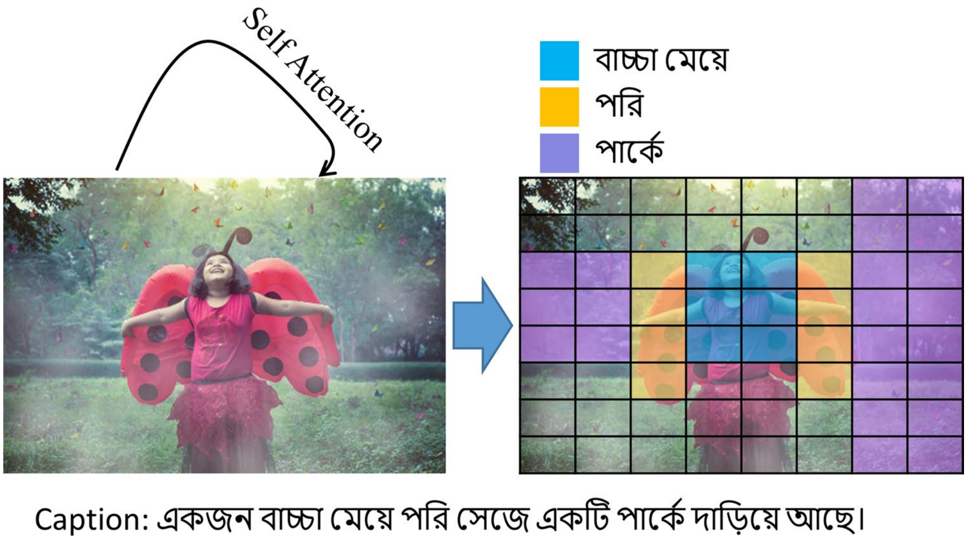

Self Attention
বাচ্চা মেয়ে
পরি
পার্কে
Caption: একজন বাচ্চা মেয়ে পরি সেজে একটি পার্কে দাড়িয়ে আছে।

Image Caption Generation in English

The starting research on image caption generation concentrated on two types of methods: template-based and retrieval-based methods [11]. These methods were not sufficiently described [11] and utilized a hard-coded language structure. Hence, these methods no longer expanded, and the neural network emerged in this field [12]. The next advanced approach was the attention-based encoder–decoder model. Xu et al. [13] presented the first attention model for generating image captions. This attention model added randomized image weights. Therefore, some necessary segments of the image were missed to generate image descriptions. You et al. [6] presented a semantic attention framework to overcome this weakness. It concentrated on the linguistically relevant objects and behaviours in the image. After that, Lu et al. [14] proposed an adaptive attention-based advanced model. It automatically determined whether to utilize visual-signal or statistical-language model. It decided which image section to use if it employed the visual-signal. Chen et al. [15] proposed a new model through a model of semantic attention. They compared it with pre-trained CNN, like VGG-16 or ResNet50. Their experiments indicated that ResNet-50 gave more promising results than VGG-16. Anderson et al. [16] presented an advanced attention model for generating image captions. This model created natural captions and obtained the highest accuracy from the MS-COCO dataset.

Image Caption Generation in Regional Languages

[17] created a dataset for image captioning *'YJ captions 26k'* in Japanese. In their study, they used three learning approaches and discovered that transfer learning is the most effective method for creating Japanese image descriptions. [18] developed a caption dataset in Japanese *'STAIR Caption'* from the images of MS-COCO dataset. In Japanese language, the researchers collected image captions from two sources: machine-translator and crowdsource [19]. [20] fetched image descriptions from machine translator, crowdsource, and human translator to generate the Flicker-8k Chinese dataset. [21] created caption datasets from crowdsourcing. Excepting crowdsource captions, [22] used machine-translated sentences. [23] proposed a new model and trained it with machine-translated Chinese image captions.

In addition to captioning images in Japanese and Chinese, a little research on German, Dutch, French, and Spanish was also found in the literature. [24] proposed an image caption generation model. This model was a multilingual captioning model in which German and English image descriptions were generated in parallel using image features. According to [25], human evaluation is the prime evaluation approach for generating multilingual German-English image captions. Hence, their study utilized manual evaluation to determine the system's accuracy for generating captions. However, due to a time constraint, their work only used high-BLEU-scored captions for human evaluation. [26] presented their captioning research to produce image descriptions in Dutch. They used crowdsourcing to acquire Dutch image captions and integrated them with the Multi30k dataset. Their work examined image descriptions in Dutch with English and French based on Multi30k dataset and noticed distinct captions for distinct languages reasons for their cultural dissimilarity. [10] presented a system for visually impaired people to generate and verbalize Spanish image descriptions.

TextMage is an image captioning method presented by [27] that generates Bengali image captions with a South Asian language bias. [28] presented a deep learning model for image captioning in Bengali. It employed a one-dimensional convolutional neural network (CNN) using the ResNet-50 pre-trained model for fetching visual features. Researchers in this [29] paper employed a hybrid technique on InceptionResNetV2 or Xception as CNN and Bidirectional LSTM or RNN on two Bengali datasets. [30] developed a multi-tasking and pre-trained word embeddings algorithm for generating Arabic language captioning.

Our Observation from the Earlier Literature

We found numerous approaches in the literature for captioning images in English and other regional languages. But we did not locate any high-quality research on image captions in Bengali that gave satisfactory results. After analyzing the past approaches, we have found that the existing datasets are erroneous as they majorly depend on Google Translate. We have cured the datasets using native speakers and decided to consider a neural network model with visual attention for generating captions in Bengali. Here, we have used curated publicly available datasets *'BanglaLekhaImageCaptions'* [31] and *'Bangla Natural Language Image to Text (BNLIT)* [32] to train our model.

Proposed Method

Here, we present our proposed model in detail. The model is composed of an encoder and a decoder. One CNN-based encoder with an attention module and one RNN-based decoder are the main component of the method. The method depicted in Fig. 2. Firstly, an input RGB image passes

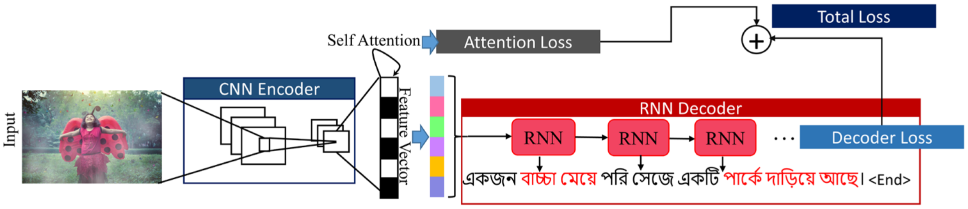


Fig. 2 The modules of the proposed system. The caption (A baby girl is standing in a park dressed as a angel). We introduced a joint learning by minimizing the total loss obtained from the visual attention loss and the decoder loss

through a CNN model. The model converts the input image into a one-dimensional vector. The CNN is pre-trained in imagenet dataset and used as a feature extractor. Next, a visual attention mechanism is applied to highlight the important areas of the image. Finally, an RNN-based decoder is used to interpret the caption, and a joint loss function is used to learn the caption generation process. Next, we discuss each module in detail.

CNN-based Encoder: The first goal of the method is to extract a low dimension feature from the input images. The CNN module in the model can be any baseline CNN. We used Inception V3 [33], and it can be replaced by any superior. The CNN model is pre-trained on Imagenet [34]. The method considers 299 × 299 RGB image as input and produces 2048 × 1 vector known as a feature vector. The weights of the final dense layer are the visual features of the input image, say $f_t = CNN - Enoder(input)$. The t denotes the temporal position of the word w_i.

RNN-based Decoder: A Gated Recurrent unit (GRUs) [35] is used as a decoder. First, a 256 dimension word embedding is used to convert the caption into a 256 × 1 vector, followed by two fully connected layers that combine the word embedding and the maximum learning capabilities in the model (we have used 512 as the default settings). It is a classification problem. We have the words as the classes. The main difference is that we take this problem as a temporal classification problem where the sequence matters. Loss is defined by the alignment matching between the word (w_i) and the feature (f_t).

Joint Loss: We use a joint loss extracted from CNN and the RNN. The attention loss ($L_{attention}$) is extracted from a pre-trained CNN and RNN both. First, the pre-trained CNN encoder converts the input image to a 2048 × 1 vector. Next, the RNN learns to focus (attention) on the next predicted word (w_i). The attention is defined by:

$$a(w_t) = v^T tanh(W_1 * w_{t-1} + W_2 * w_{t+2}) \tag{1}$$

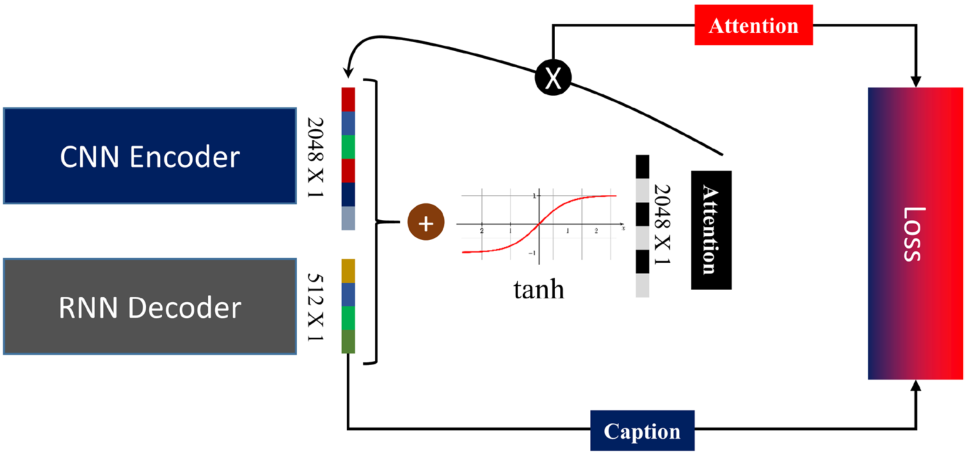
Where v is the weight vector, W_1, W_2 are word weight.

The attention mechanism uses the time-space relation between time (t) and spatial feature (f_t). The words (w_i) showed up in such time (t) are associated with the structure of the description. The recurrent visual loss is estimated as:

$$L_{visual} = 1 - \sum_{j=1}^{N}(match(a(w_t), f_t)) \tag{2}$$

Next, the RNN decoder combines the feature vector and the sequence. It is represented using a context vector (κ). The RNN maps the captions with visual attention, considering the sequence or time. $L_{sequence}$ is defined by:

Fig. 3 Loss contains a combination of spatial loss and a temporal loss. CNN encoder provides the spatial or visual attention, and RNN decoder provides the alignment of the visual attention with the caption sentence



$$L_{sequence}(\kappa) = -\sum_{i=1}^{m}\sum_{t=1}^{n}\log y_{it} \tag{3}$$

The complete method for calculating the proposed loss is depicted in Fig. 3

Parameters Settings: An input of 299 × 299 RGB image is taken. Next, the image passes through the baseline CNN model for extracting a 2048 × 1 feature vector. Next, a tanh layer is used to extract 2048 × 1 visual attention. The attention followed a multi-resolution attention procedure. The different heads of the attention are the skip connection among different convolution modules of the Inception network. The major parameters of the RNN are the hidden units (512), the attention map (2048 × 1), and the length of words embedding (256). Here, we have used joint loss combined with Adam optimizer for the joint training (CNN and RNN). We have used 120 epochs for the training.

Dataset

We have used two datasets *Bangla Lekha Image Captions (BLIC)* [31] and *Bangla Natural Language Image to Text (BNLIT)* to train the captioning model in Bengali. The BLIC dataset has 9,154 images with two captions per image, and the BNLIT dataset includes 8,743 images with one caption per image. Both the datasets are smaller than the available English datasets. The datasets retain some relevance towards Bengali culture. Human bias is prevalent in the datasets. This bias prevents the ability of any model to create captions for non-human subjects. Sometimes, these captions are not in detail enough. As a result, the training of any model and the evaluation accuracy utilizing these datasets are not as expected. Each dataset was separated into three parts, 80% of the images utilized for training, 10% for validating, and the remaining 10% for testing.

Table 1 Accuracy of the proposed method using BLIC & BNLIT datasets

Method	BLIC		BNLIT	
	BLEU-1	BLEU-4	BLEU-1	BLEU-4
Mixture Model [39]	0.63	0.16	0.61	0.14
Injection Model [40]	0.61	0.16	0.58	0.15
TextMage [27]	0.65	0.23	–	–
Encoder-decoder [28]	0.58	0.17	0.54	0.16
Xception+BiGRU [29]	0.67	0.24	–	–
Proposed+VGG16 [37]	0.52	0.16	0.50	0.14
Proposed+ResNet50 [38]	0.58	0.18	0.55	0.17
Proposed+Inception [33]	**0.67**	**0.25**	**0.65**	**0.24**

Bold values indicate the best results

Evaluation Metric

The BLEU evaluation metric is used to test the accuracy of our experiment. In 2002, Papineni et al. proposed the BLEU as a popular evaluation metric [36]. It is used a lot in NLP and computer vision applications like machine translation, image captioning, etc. The number of n-grams in the generated sentence is compared to the number of n-grams in the reference sentence to determine the BLEU score. The different BLEU scores are measured based on the number of word sequences in n-gram. For the unigram-similarity BELU-1, for bigram BLEU-2, for trigram BLEU-3 is used, and so on.

Experimental Results with Discussion

The BLEU evaluation measure was used to assess the image captions generated by our proposed approach. Table 1 depicts a comparison between the proposed model and existing models. We have tested our model with the VGG16 [37], ResNet-50 [38] and Inception V3 [33] model and found Inception V3 outperformed the other models (Table 2).

This research proposes a method for creating Bengali image captions using the visual attention model. We have observed our visual attention model outperforms other models in Bengali datasets. We have tested our proposed model with 120 epochs, and the loss is very low to the end at 0.002. It denotes the model stability of the proposed model. Figure 4 depicts the loss curve varying epochs during training. Figure 5 depicts the loss curve varying epochs during training. Figure 6 shows the attention regions to generate captions from three images. Figures 7 and 8 visualize some sample results of our proposed model.

Previously, we have tested our method varying by different CNN and RNN baseline modules. We have also performed another ablation study by removing attention loss and decoder loss. We have observed that the method outperforms when we use both of the losses. The study is shown in Table 3.

Table 2 Accuracy of the proposed method on BLIC & BNLIT datasets using cross-dataset training

Method	BLIC		BNLIT	
	BLEU-1	BLEU-4	BLEU-1	BLEU-4
Proposed+VGG16 [37]	0.31	0.11	0.22	0.09
Proposed+ResNet50 [38]	0.42	0.14	0.42	0.12
Proposed+Inception [33]	**0.52**	**0.14**	**0.42**	**0.16**

Bold values indicate the best results

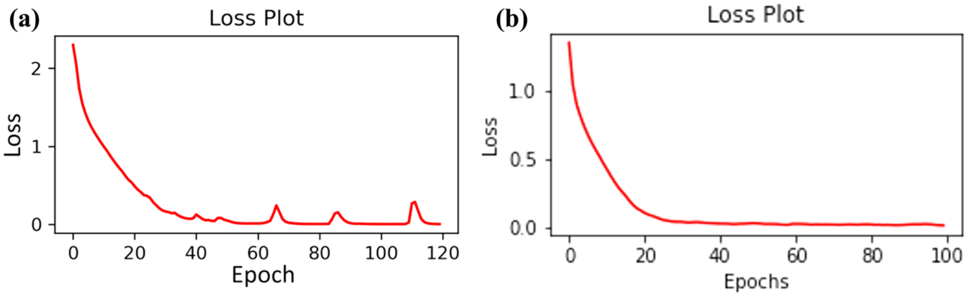


Fig. 4 Loss over epochs during training. **a** Loss curve (BLIC dataset), **b** Loss curve (BNLIT dataset)

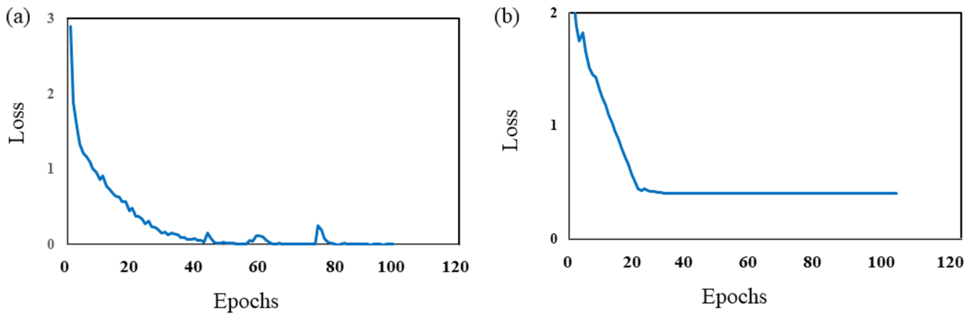
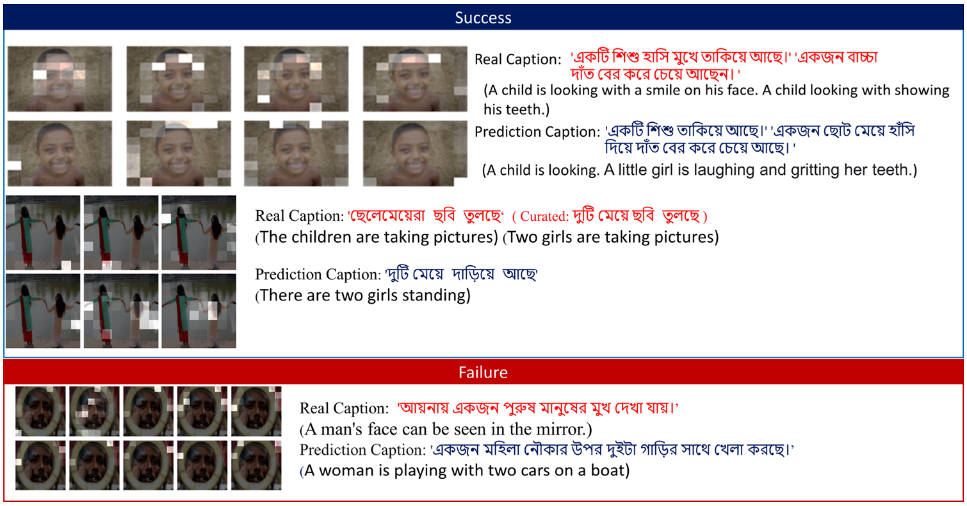


Fig. 5 Validation Loss over epochs during training. **a** Validation Loss curve (BLIC dataset), (**b**) Validation Loss curve (BNLIT dataset)



Real Caption: 'একটি শিশু হাসি মুখে তাকিয়ে আছে।' 'একজন বাচ্চা দাঁত বের করে চেয়ে আছেন।'
(A child is looking with a smile on his face. A child looking with showing his teeth.)
Prediction Caption: 'একটি শিশু তাকিয়ে আছে।' 'একজন ছোট মেয়ে হাঁসি দিয়ে দাঁত বের করে চেয়ে আছে।'
(A child is looking. A little girl is laughing and gritting her teeth.)

Real Caption: 'ছেলেমেয়েরা ছবি তুলছে' (Curated: দুটি মেয়ে ছবি তুলছে)
(The children are taking pictures) (Two girls are taking pictures)

Prediction Caption: 'দুটি মেয়ে দাড়িয়ে আছে'
(There are two girls standing)

Real Caption: 'আয়নায় একজন পুরুষ মানুষের মুখ দেখা যায়।'
(A man's face can be seen in the mirror.)
Prediction Caption: 'একজন মহিলা নৌকার উপর দুইটা গাড়ির সাথে খেলা করছে।'
(A woman is playing with two cars on a boat)

Fig. 6 Examples of various visual attention regions generated by the proposed method along with captions. The blue horizontal bar denotes two successful cases from (BLIC) [31] and (BNLIT) datsets. The Magenta horizontal bar is a failure case from BNLIT dataset. It is also shown that the success example from BLNIT dataset contains erroneous groundtruth, we have curated that

Conclusion

This article presents a deep learning model for generating image captions in Bengali. The proposed model uses the visual attention mechanism to generate image captions. Two datasets BLIC and BNLIT were used to train the model. The BLIC dataset contains two descriptions per image, and the BNLIT dataset contains one description per image. Our model achieved higher BLEU-1 and BLEU-4 scores than other models. To increase the BLEU score, our objective is to train the model with additional captions per image in the future. This model has trouble to recognize non-human subjects since the dataset is biased towards human subjects. Therefore, developing a detailed and well-varied Bengali captioning dataset is necessary to enhance accuracy. However, the overall result shows that this model potentially contributes to regional image-captioning research and also be further used for generating questions from an image in regional languages.

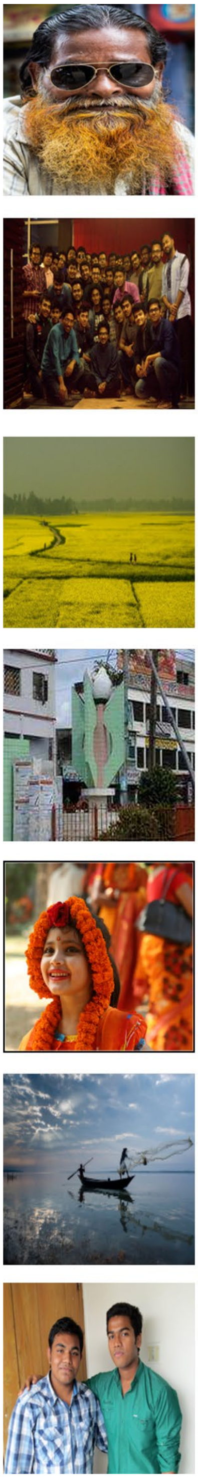


Fig. 7 The random images from the BLIC dataset are taken to visualize the result. R indicates the original or true captions, and P denotes the predicted or system-generated captions

R: একজন বয়স্ক পুরুষ আছে। একজন বুড়ো পুরুষ কালো সান গ্লাস পরে তাকিয়ে আছেন যার চুল কালো দাড়িগোঁফ হলুদ সাদা।
There is an old man. An old man is wearing black sunglasses with black hair, yellow and white beard.

P: চশমা পরিহিত একজন বয়স্ক পুরুষ তাকিয়ে আছে। একজন বয়স্ক পুরুষ তাকিয়ে আছে।
An elderly man wearing glasses is staring. An old man is staring.

R: একসাথে অনেকগুলো পুরুষ আছে। কিছু ছেলে একসাথে দাড়িয়ে এবং বসে আছে ছবি তোলার জন্য।
There are many men together. Some boys are standing and sitting together to take pictures.

P: কয়েকজন মানুষ বসে আছে আর পিছনে অনেক মানুষ দাঁড়িয়ে আছে।, কয়েকজন ছেলে দাড়িয়ে আছে।
A few people are sitting and many people are standing behind. A few boys are standing.

R: সরিষা ক্ষেতের মাঝ দিয়ে দুইজন মানুষ হেঁটে যাচ্ছে।, অনেকগুলো সরিষা ক্ষেত যার মাঝ দিয়ে একটি রাস্তা দিয়ে হেটে যাচ্ছে ৩ জন মানুষ।
Two people are walking through the middle of a mustard field. There are many mustard fields through which 3 people are walking along a road.

P: সরিষা ক্ষেতের মাঝ দিয়ে দুইজন মানুষ হেঁটে যাচ্ছে।, সরিষার ক্ষেত দিয়ে তিন জন যাচ্ছে।
Two people are walking through the mustard field. Three people are walking through the mustard field.

R: ভবন দেখা যাচ্ছে।, একটি তোরণ দেখা যাচ্ছে।
There is a building. There is a gate.

P: এটি একটি ভবন আছে। পিছনে দূরে একটি বিল্ডিং স্কুল।
It is a building. A building school in the back.

R: একটি শিশু ফুল দিয়ে সেজেছে।, একটি বাচ্চা মেয়ে ফুল মাথাই দিয়ে হাসছে।
A child is dressed with flowers, A baby girl is smiling with a flower head.

P: একটি শিশু তাকিয়ে আছে।, একজন ছোট মেয়ে হাঁসি দিয়ে দাঁত বের করে চেয়ে আছে।
A child is looking. A little girl is laughing with gritting her teeth.

R: নৌকায় দুইজন মানুষ আছে।, ১ টি নৌকায় বসে আছে মাঝি এবং আরেকজন পুরুষ নৌকা থেকে জাল ফেলছে পানিতে।
There are two people in the boat, A boatman is sitting in one boat and another man is throwing nets from the boat into the water.

P: দুইজন মানুষ একটি নৌকা চালাচ্ছে।, নদীর উপর দিয়ে মাছ ধরছে।
Two people are driving a boat, Fishing on the river.

R: পাশাপাশি দুইজন পুরুষ আছে।, ২ জন ছেলে একজন আরেকজনের ঘাড়ে হাত রেখে দাড়িয়ে আছে।
There are two men side by side. Two boys are standing with their hands on each other's necks.

P: দুইজন পুরুষ দাড়িয়ে আছে।, ২ জন ছেলে দাড়িয়ে আছে।
Two men are standing. There are 2 boys standing.

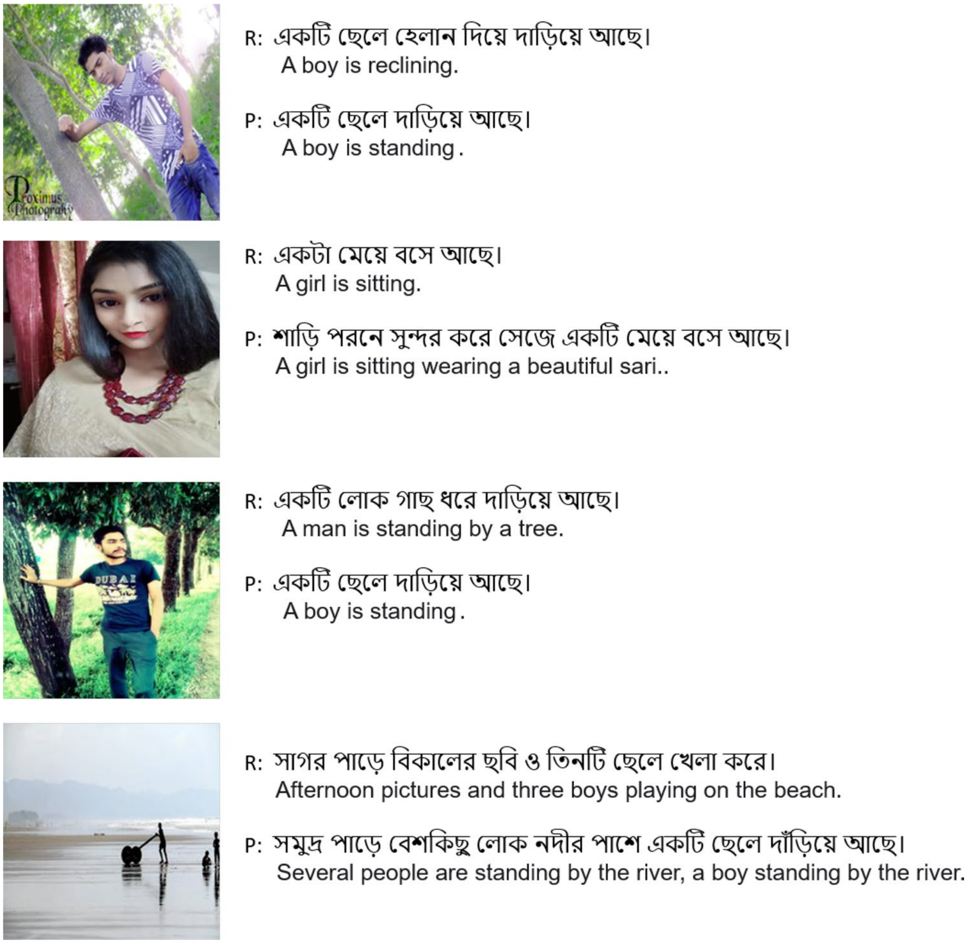

Proximus Photography
R: একটি ছেলে হেলান দিয়ে দাড়িয়ে আছে।
A boy is reclining.
P: একটি ছেলে দাড়িয়ে আছে।
A boy is standing .
R: একটা মেয়ে বসে আছে।
A girl is sitting.
P: শাড়ি পরনে সুন্দর করে সেজে একটি মেয়ে বসে আছে।
A girl is sitting wearing a beautiful sari..
R: একটি লোক গাছ ধরে দাড়িয়ে আছে।
A man is standing by a tree.
DUBAI
P: একটি ছেলে দাড়িয়ে আছে।
A boy is standing .
R: সাগর পাড়ে বিকালের ছবি ও তিনটি ছেলে খেলা করে।
Afternoon pictures and three boys playing on the beach.
P: সমুদ্র পাড়ে বেশকিছু লোক নদীর পাশে একটি ছেলে দাঁড়িয়ে আছে।
Several people are standing by the river, a boy standing by the river.

Fig. 8 The result of the proposed model is visualized using random images from the BNLIT dataset

Table 3 Accuracy of the proposed method on BLIC & BNLIT datasets using different losses

Method	BLIC		BNLIT	
	BLEU-1	BLEU-4	BLEU-1	BLEU-4
Attention loss	0.22	0.09	0.16	0.11
Decoder loss	0.19	0.11	0.26	0.14

Data availability The source code and the dataset are available at https://github.com/bidyut2002in/A-Visual-Attention-Based-Model-for-Bengali-Image-Captioning.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Consent for publication All authors read and approved the final manuscript.

References

1. Mishra SK, Dhir R, Saha S, Bhattacharyya P, Singh AK. Image captioning in hindi language using transformer networks. Computers & Electrical Engineering. 2021;92: 107114.
2. Das B, Sekh AA, Majumder M, Phadikar S, Abid: Attention-based bengali image description. In: Proceedings of the 3rd International Conference on Communication, Devices and Computing, 2022;305–314 . Springer
3. Vinyals O, Toshev A, Bengio S, Erhan, D, Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015;3156–3164
4. Johnson J, Karpathy A, Fei-Fei L, Densecap: Fully convolutional localization networks for dense captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016;4565–4574
5. Karpathy A, Fei-Fei L, Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015;3128–3137
6. You Q, Jin H, Wang Z, Fang C, Luo J, Image captioning with semantic attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016;4651–4659
7. Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research. 2013;47:853–99.

8. Young P, Lai A, Hodosh M, Hockenmaier J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics. 2014;2:67–78.
9. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL, Microsoft coco: Common objects in context. In: European Conference on Computer Vision, 2014;740–755. Springer
10. Gomez-Garay A, Raducanu B, Salas J, Dense captioning of natural scenes in spanish. In: Mexican Conference on Pattern Recognition, 2018;145–154 . Springer
11. Bai S, An S. A survey on automatic image caption generation. Neurocomputing. 2018;311:291–304.
12. Hossain MZ, Sohel F, Shiratuddin MF, Laga H. A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR). 2019;51(6):1–36.
13. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y, Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning, 2015;2048–2057 . PMLR
14. Lu J, Xiong C, Parikh D, Socher R, Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017;375–383
15. Chen Q, Li W, Lei Y, Liu X, He Y, Learning to adapt credible knowledge in cross-lingual sentiment analysis. In: Proceedings of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on NLP (Volume 1: Long Papers), 2015;419–429
16. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L, Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018;6077–6086
17. Miyazaki T, Shimizu N, Cross-lingual image caption generation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016;1780–1790
18. Yoshikawa Y, Shigeto Y, Takeuchi A, Stair captions: Constructing a large-scale japanese image caption dataset. 2017, arXiv preprint arXiv:1705.00823
19. Rathi A, Deep learning apporach for image captioning in hindi language. In: 2020 International Conference on Computer, Electrical & Communication Engineering (ICCECE), 2020;1–8 . IEEE
20. Li X, Lan W, Dong J, Liu H, Adding chinese captions to images. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, 2016;271–275
21. Li X, Xu C, Wang X, Lan W, Jia Z, Yang G, Xu J. Coco-cn for cross-lingual image tagging, captioning, and retrieval. IEEE Trans Multimedia. 2019;21(9):2347–60.
22. Lan W, Li X, Dong J, Fluency-guided cross-lingual image captioning. In: Proceedings of the 25th ACM International Conference on Multimedia, 2017;1549–1557
23. Zeng X, Wang X, Add english to image chinese captioning. In: 2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2017;333–338 . IEEE
24. Elliott D, Frank S, Hasler E, Multilingual image description with neural sequence models. 2015, arXiv preprint arXiv:1510.04709
25. Elliott D, Frank S, Sima'an K, Specia L, Multi30k: Multilingual english-german image descriptions.2016, arXiv preprint arXiv:1605.00459
26. van Miltenburg E, Elliott D, Vossen P, Cross-linguistic differences and similarities in image descriptions. 2017, arXiv preprint arXiv:1707.01736
27. Kamal AH, Jishan MA, Mansoor N, Textmage: The automated bangla caption generator based on deep learning. In: 2020 International Conference on Decision Aid Sciences and Application (DASA), 2020;822–826 . IEEE
28. Khan MF, Shifath S, Islam M, et al. Improved bengali image captioning via deep convolutional neural network based encoder-decoder model. 2021, arXiv preprint arXiv:2102.07192
29. Humaira M, Paul S, Jim M, Ami AS, Shah FM. A hybridized deep learning method for bengali image captioning. IJACSA. 2021;12(2):698–707.
30. Eddin Za'ter M, Talaftha B, Bench-marking and improving arabic automatic image captioning through the use of multi-task learning paradigm. arXiv e-prints, 2202 (2022)
31. Mansoor N, Kamal AH, Mohammed N, Momen S, Rahman MM, Banglalekhaimagecaptions, mendeley data 2019. (Date last accessed 15-July-2014). http://dx.doi.org/10.17632/rxxch9vw59.2
32. Jishan MA, Mahmud KR, Al Azad AK, Ahmmad MR, Rashid BP, Alam MS. Bangla language textual image description by hybrid neural network model. Indonesian Journal of Electrical Engineering and Computer Science. 2021;21(2):757–67.
33. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z, Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016;2818–2826
34. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, et al. Imagenet large scale visual recognition challenge. Int J Comput Vision. 2015;115(3):211–52.
35. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y, Learning phrase representations using rnn encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014;1724–1734
36. Papineni K., Roukos S, Ward T, Zhu W-J, Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002;311–318
37. Simonyan K, Zisserman A, Very deep convolutional networks for large-scale image recognition. 2014, arXiv preprint arXiv:1409.1556
38. He K, Zhang X, Ren S, Sun J, Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016;770–778
39. Tanti M, Gatt A, Camilleri K, What is the role of recurrent neural networks (rnns) in an image caption generator? In: Proceedings of the 10th International Conference on Natural Language Generation, 2017;51–60
40. Rahman M, Mohammed N, Mansoor N, Momen S. Chittron: An automatic bangla image captioning system. Procedia Computer Science. 2019;154:636–42.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.