CapNet: An Encoder-Decoder based Neural Network Model for Automatic Bangla Image Caption Generation

Rashik Rahman[1]
Computer Science and Engineering
University of Asia Pacific,
Dhaka, Bangladesh

Hasan Murad[2]
Computer Science and Engineering
Chittagong University of
Engineering and Technology
Chattogram, Bangladesh

Nakiba Nuren Rahman[3]
Computer Science and Engineering
University of Asia Pacific,
Dhaka, Bangladesh

Aloke Kumar Saha[4]
Computer Science and Engineering
University of Asia Pacific,
Dhaka, Bangladesh

Shah Murtaza Rashid Al Masud[5]
Computer Science and Engineering
University of Asia Pacific,
Dhaka, Bangladesh

A S Zaforullah Momtaz[6]
Computer Science and Engineering
University of Asia Pacific,
Dhaka, Bangladesh

***Abstract*—Automatic caption generation from images has become an active research topic in the field of Computer Vision (CV) and Natural Language Processing (NLP). Machine generated image caption plays a vital role for the visually impaired people by converting the caption to speech to have a better understanding of their surrounding. Though significant amount of research has been conducted for automatic caption generation in other languages, far too little effort has been devoted to Bangla image caption generation. In this paper, we propose an encoder-decoder based model which takes an image as input and generates the corresponding Bangla caption as output. The encoder network consists of a pretrained image feature extractor called ResNet-50, while the decoder network consists of Bidirectional LSTMs for caption generation. The model has been trained and evaluated using a Bangla image captioning dataset named BanglaLekhaImageCaptions. The proposed model achieved a training accuracy of 91% and BLEU-1, BLEU-2, BLEU-3, BLEU-4 scores of 0.81, 0.67, 0.57, and 0.51 respectively. Moreover, a comparative study for different pretrained feature extractors such as VGG-16 and Xception is presented. Finally, the proposed model has been deployed on an embedded device for analysing the inference time and power consumption.**

Keywords*—*Bangla image caption generation; encoder-decoder; bidirectional long short term memory (LSTM); bangla natural language processing (NLP)

I. Introduction

A picture is equivalent to million of stories. It is simple for people to narrate these stories, but challenging for machines to illustrate them. In the domain of intuitive systems, machine generated image captioning is an amalgamation of computer vision and NLP. Semantically and syntactically correct image caption generation is challenging for the machine compared to human beings. However, automatic caption generation from image content has a significant number of real life applications from the field of human machine interaction (HCI) to robotics.

According to World Health Organization (WHO), almost 2.2 billion people in the world have a near or distance vision impairment[1]. Automatic image caption generation plays a significant role for visually impaired people by converting the caption to speech to have a better understanding of their surroundings.

Automatic speech generation for the humanoid robot is a challenging task which involves generating caption by understanding the robot vision. Therefore, automatic image caption generation has considerable impact in the field of robotics. Content creation for social media platforms has become a professional sector which has created a large job sector for the young generation. However, content needs proper captioning before publishing in social media platforms. Therefore, providing automatic suggestions for image captioning is handy for content creators on social media platforms.

In recent years, image caption generation has become a relatively active field of research and therefore a significant number of research has been found in literature where most of the researchers focus on image caption generation in the English language [1], [2].

Though Bangla is the seventh largest language in the world with 215 million speakers globally[2], far too little effort has been devoted to Bangla image caption generation. Researchers have not addressed automatic image captioning in Bangla for a long period of time due to a lack of an enriched dataset. After development of required dataset, several researches have been conducted on Bangla caption generation from visual image [3], [4], [5], [6], [7].

However, the performance metrics given in the previous related work show that the quality of the generated Bangla image caption is not quite satisfactory. Therefore, there is a clear scope for further improvements in automatic Bangla image captioning. Moreover, we did not find any attempt to deploy the model on an embedded device.

[1] shorturl.at/hRWZ6

[2] $https : //www.vistawide.com/languages/top_30_languages.htm$

The objective of this research work is to develop a image captioning model that can automatically generate Bangla caption with better performance compared to the models found in the previous related works. In addition to proposing an end-to-end system, the trained captioning model is deployed within an embedded device in order to evaluate the efficiency of the model.

We designed a model architecture using a deep learning based encoder-decoder model which takes an image as input and generates the corresponding Bangla caption as output. The encoder network consists of a pretrained image feature extractor while Bidirectional LSTMs are used in the decoder network for caption generation. We explore different pretrained image feature extractors such as VGG-16, Xception, and ResNet-50 for the encoder network. The model is trained and evaluated using a Bangla image captioning dataset named BanglaLekhaImageCaptions. We have achieved the best training results for the encoder-decoder model with the ResNet-50 pretrained feature extractor. The final training accuracy during convergence is 91% and the BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores are 0.81, 0.67, 0.57, and 0.51, respectively, which is the state of the art result for automatic Bangla image captioning.

The major contributions of this paper are:

- Trained and validated the state of the art model for automatic Bangla image caption generation using an encoder-decoder based model architecture.
- Explored different pretrained image feature extractors such as VGG-16, Xception, and ResNet-50 for the encoder network and found the model with ResNet-50 provides the best BLEU score for Bangla image captioning.
- Finally, Deployed the proposed model on an embedded device for analysing the key performance metrics such as the inference time and power consumption.

The remaining sections of the paper are organized as follows: The literature review covered in the Section II. The Section III presents an overview of the dataset. Section IV provides a comprehensive breakdown of the proposed system. In Section V, all the experimental details during training and validation of our model are stated. In Section VI, findings and comparisons of this research is provided. The conclusion is located in the Section VII. The remainder consists of references.

II. Literature Review

In this section, we illustrate the evolution of research in the field of automatic caption generation for images. Moreover, the recent development of Bangla image captioning in literature is presented.

After conducting a rigorous literature review, we have found that there are two types of techniques based on traditional machine learning and deep learning to generate automatic image captioning [8], [9].

Early research on automatic image captioning utilises traditional machine learning techniques such as similar image retrieval based captioning [10] and template matching based image captioning [11]. However, the generated captions are limited by a predefined corpus with images or templates and their corresponding captions as labels. Therefore, traditional machine learning techniques fail to generate relevant image captions if the input image has significant differences from the predefined corpus.

The recent impressive progress in the fields of computer vision and NLP has paved the way to deploying deep learning techniques to generate image captions automatically. Image captioning involves vision encoding for a high-level understanding of image features and language decoding for caption generation using the features generated from vision encoding. The encoder-decoder based deep learning model is the most effective technique to address vision encoding and language decoding. In literature, vision encoder is designed using stacked Convolutional Neutral Network (CNN) [1], and graph-based network [2]. Moreover, various pre-trained feature extractors such as VGG-16, InceptionResnetV2, and Xception have been deployed for vision encoding [3], [12] The language decoder is implemented using variations of Recurrent Neural Networks (RNNs) such as LSTMs and GRUs [2]. In addition, self attention based transformer models are utilised to design the language decoder [13].

A considerable amount of effort has been devoted to developing automatic image captioning techniques in languages such as English [1], Chinese [14], Japanese [15], Arabic [16], Hindi [17] and German [18] where large datasets related to image captioning are already available.

Due to a lack of an enriched dataset, researchers have not addressed automatic image captioning in Bangla for a long period of time. However, after the development of the required dataset, Bangla image captioning has become an active research area among researchers. Table I presents an overview of related literature on Bangla image captioning with information on model architecture designed, dataset used during training and evaluation, and BLEU score as evaluation metrics to measure the quality of the generated caption by the model.

Rahman et al. [12] has developed the first Bangla image captioning dataset named BanglaLekhaImageCaptions. They have trained and evaluated an encoder-docoder model using their own dataset, where the encoder network utilises a pre-trained feature extractor called VGG-16, and the decoder network is designed using stacked LSTMs network. However, they have not calculated the BLEU score on their whole test dataset and have only reported the BLEU score for a few sample test images during evaluation, where the BLEU score is unsatisfactory.

Kamal et al. [6] have proposed a similar encoder-decoder model mentioned in [12] for Bangla image captioning where the encoder network consists of a VGG-16 pre-trained model and the decoder network consists of LSTMs network. Moreover, they have utilised the same BanglaLekhaImageCaptions dataset for training the model. However, they have evaluated the model by calculating the BLEU score for the test dataset, which was missing in [12]. The achieved BLEU-1 score for the model is 0.67 on the test dataset.

Jishan et al. [4] have proposed a hybrid encoder-decoder

TABLE I. An Overview of Recent Research Works on Bangla Image Caption Generation

Research	Year	Dataset	Modeling techniques	Performance
Humaira et al. [3]	2021	BanglaLekhaImageCaption	InceptionResnetV2 or Xception + BiLSTM or BiGRU	BLEU-1: 0.674, BLEU-2: 0.53, BLEU-3: 0.45, BLEU-4: 0.344
Khan et al. [5]	2021	BanglaLekhaImageCaption	1D CNN+ResNet-50	BLEU-1: 0.65, BLEU-2: 0.45, BLEU-3: 0.28, BLEU-4: 0.175
Palash et al. [7]	2021	BanglaLekhaImageCaption	ResNet-101+Attention mechanism+decoder	BLEU-1: 0.69, BLEU-2: 0.63, BLEU-3: 0.58
Kamal et al. [6]	2020	BanglaLekhaImageCaption	VGG-16+LSTM	BLEU-1: 0.67, BLEU-2: 0.44, BLEU-3: 0.32, BLEU-4: 0.24
Jishan et al. [4]	2020	BNLIT	CNN+BiLSTM	BLEU-1: 0.65, BLEU-2: 0.47, BLEU-3: 0.33, BLEU-4: 0.23

based model where they suggested a custom CNN architecture responsible for extracting image features and utilising Bidirectional Long Short Term Memory (BiLSTM) as a decoder for caption generation. They have trained and evaluated their model using their own dataset called Bangla natural language image to text (BNLIT). They have achieved a BLEU-1 score of 0.65 after evaluating their model on the test dataset.

Khan et al. [5] have suggested an end-to-end image captioning system where ResNet-50 is for image feature extraction and one dimensional CNN for generating captions, and they used BanglaLekhaImageCaption dataset to train and test their model. Their proposed system achieved a BLEU-1 score of 0.65.

Humaira et al.[3] have presented a performance evaluation of Bangla captioning systems using pre-trained models such as InceptionResnetV2, Xception as encoders and BiLSTM or BiGRU as decoders while using the BanglaLekhaImageCaptions dataset. They have achieved a maximum BLEU-1 score of 0.674 after evaluating their model.

Palash et al. [7] have provided a novel transformer-based architecture that automatically generates Bangla captions from an input image. They have proposed a new transformer architecture with an attention mechanism as a decoder and employed ResNet-101 as an encoder. They have trained and evaluated the model using the BanglaLekhaImageCaptions dataset and achieved a BLEU-1 score of 0.69.

From the previous related works, it is evident that there is a clear scope for further improvement in BLEU score of the Bangla image captioning model. In addition, we did not find any attempt to deploy the model on an embedded device.

In this research work, an encoder-decoder network for Bangla image caption generation and explore different pre-trained image feature extractors for the encoder network is proposed. Finally, we deploy the Bangla image captioning model with the best BLEU score onto an embedded device.

III. Dataset

In the research work, the BanglaLekhaImageCaptions dataset proposed by Rahman et al. [12] is utilized. The downloadable dataset is available online in Mendeley Data[3]. This dataset includes photos with Bengali annotations.

All of its captions are annotated by native Bengali people. There are only two captions tagged with each image in this dataset, yielding a total of 18308 descriptions for the 9154 images. BanglaLekha has 5270 distinct Bengali words. All popular picture captioning datasets are primarily influenced by western culture, with the majority of annotations performed in English. Using such datasets to train an image captioning system for Bangla is not effective. Thus, requiring the necessity for a culturally significant dataset in Bengali to generate acceptable image captions from images related to Bangladeshi and greater sub-continental culture. From the dataset, 80% data is used to train the model, and after training, the remaining 20% is used to evaluate and validate the model.

IV. Proposed System

In this section, you present the proposed model architecture for Bangla image captioning.

We have designed an encoder-decoder based model architecture. Fig. 1 shows a high level overview of the model architecture. As we are working with both image data as input and text data as output, extraction features from image and application of word embedding to the text data is required.

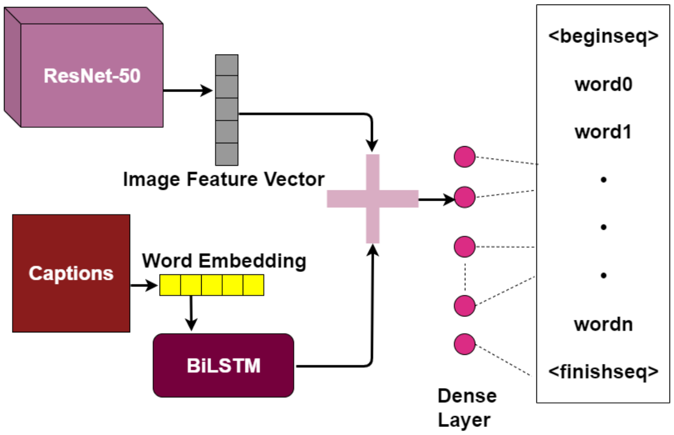


Fig. 1. Overview of Caption Generation Process.

Feature Extraction

A key component of image captioning is the extraction of visual features. We are deploying a pre-trained model called ResNet-50, which has already been trained on millions of images from the ImageNet dataset [19]. As the model is solely utilised for feature extraction, the final two layers have been eliminated, leaving the GlobalAveragePooling layer instead of

[3] https://data.mendeley.com/datasets/rxxch9vw59/2

a dense layer as the final layer. In contrast to the MaxPooling layer, which generates a 2D matrix, the GlobalAveragePooling layer generates a vector with a dimestion of (None, 2048). The input shape for ResNet-50 is (224,224,3). Therefore, all the photos are reshaped to match this dimension. In the input shape, 3 specifies the number of channels, since the images are in RGB format, the channel number is set to 3.

Word Embedding

Before passing words to RNNs like LSTMs or BiLSTMs, they must be embedded, which turns words into vectors. The embedding layer makes it possible to turn each word into a vector of fixed length and size. The generated vector is dense and contains real values as opposed to merely 0s and 1s. The fixed size of word vectors is the key reason for expressing words with fewer dimensions and in a more efficient manner. In this manner, the embedding layer functions as a lookup table, where the words are the keys and the word vectors are the values. This embedding task is accomplished using the embedding layer of the Tensorflow framework. Using an embedding layer, rather than manually setting values for each word, the embedding values are learned during training. The input and output shapes of this layer are (None, 39) and (None, 39, 128), respectively.

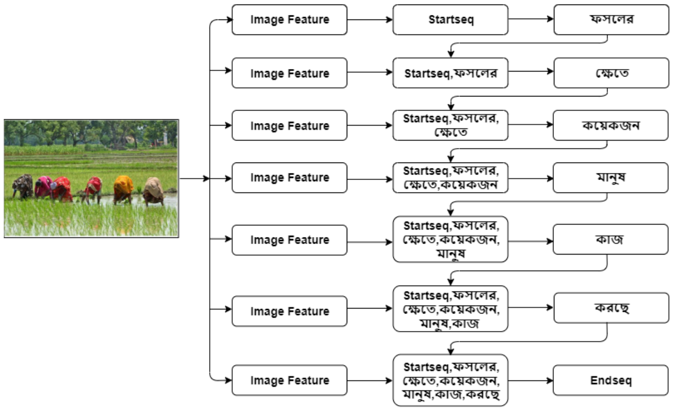


Fig. 2. Example of Word Sequence Generation.

Generation of Word Sequence

Each image in the BanglaLekhaImageCaptions dataset contains only two captions. The maximum and minimum word lengths are 39 and 26, respectively. Although a decrease in word count tends to result in a higher evaluation score [3], the goal of this research work is to develop meaningful, descriptive captions for real-life scenarios, so we use 39 as the fixed word length. During training, zero-padding is employed to increase the length of sentences that are shorter than the fixed maximum length. In addition, a beginseq token and a finishseq token are appended to each pair sequence for identification purposes throughout the training phase. In the training phase, the picture features are extracted from images and the next word in the series is generated using word vectors. Fig. 2 depicts the input-output pair.

In consideration of the limitations of RNNs, LSTMs are a superior option for word generation [20]. However, LSTMs only learn from prior words; for creating syntactically and grammatically accurate sentences, it is also necessary to preserve the knowledge of succeeding words. Therefore, the suggested model uses BiLSTMs, which retains the knowledge learned in both directions, i.e., from both preceding and succeeding words. Fig. 3 depicts the data-flow in BiLSTMs, where P_0,...,P_n are the input words and Q_0,...,Q_n are the outputs of the BiLSTMs which are determined by Eq. 1, where Q_i is output at i^{th} time when activation function h is utilized to weight W_Q and bias B_Q taking into account for forward activation m_i and backward activation n_i at i^{th} time.

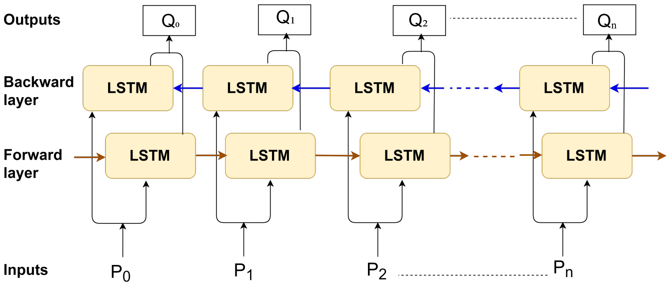


Fig. 3. Illustration of BiLSTMs having P_0,...,P_n as Inputs and Q_0,...,Q_n as Outputs.

$$Q_i = h(W_Q[m_i, n_i] + B_Q) \tag{1}$$

Each LSTM in the BiLSTMs comprises of three gates: input, forget, and output gate. The input gate indicates what incoming information will be stored in the cell state. The forget gate determines what information to discard from the cell state, whereas the output gate provides output at i^{th} time. The corresponding equations for these gates are Eq. 2, Eq. 3, Eq. 4, respectively.

$$j_i = \sigma(W_j[H_{i-1}, P_i] + B_j) \tag{2}$$

$$k_i = \sigma(W_k[H_{i-1}, P_i] + B_k) \tag{3}$$

$$l_i = \sigma(W_l[H_{i-1}, P_i] + B_l) \tag{4}$$

Here, j_i, k_i, l_i is the input, forget and output gate, sigmoid function is represented by σ, W_j, W_k, W_l are the corresponding gate's weights, H_{i-1} is considered to be previous LSTMs block's output at time $i-1$, P_i is the input at i^{th} time and B_j, B_k, B_l are the corresponding gate's bias.

Encoder

The encoder consists of two components, one for managing image feature vectors and the other for managing word sequences. ResNet-50 [21] is used to extract image features originally. These image features are transferred first to a dense layer with 128 units and then to a RepeatVector layer. The RepeatVector layer repeats the inputs for a predetermined number of times. The input to this layer is (None, 128). The RepeatVector's output shape, however, is (39, 128), as

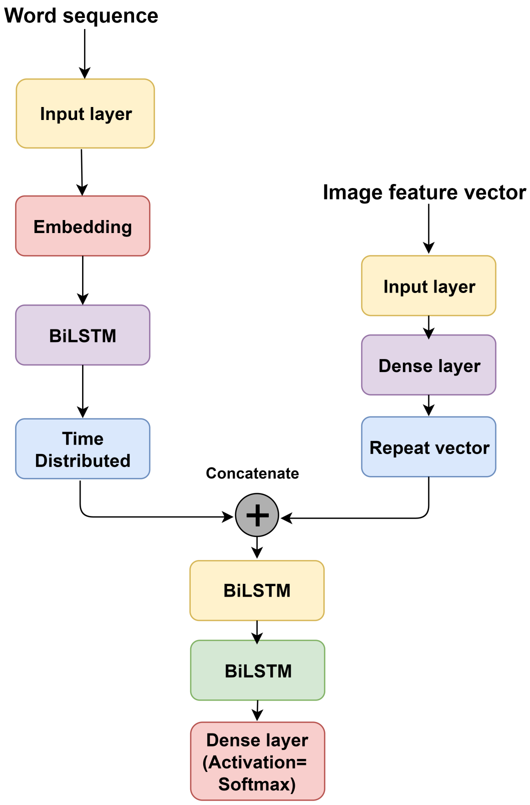


Fig. 4. Encoder-Decoder Model Architecture.

the inputs need to be repeated 39 times to match the output dimension of the other half of the encoder, which handles word vectors. The encoder's word vector processing side includes an embedding layer, a BiLSTM layer with 512 units, and finally a TimeDistributed layer. In contrast with the regular LSTMs, input travels in both directions, and knowledge from each side can be utilised in BiLSTMs. Additionally, it is a potent instrument for modelling the sequential relationships between words and phrases in each direction. In short, BiLSTMs add an additional LSTMs layer that reverses the flow of information, which indicates that the input sequence streams in reverse in the second LSTMs layer. The TimeDistributed layer uses a specified layer (a dense layer with 128 units in the suggested model) for each input vector. Both sides of the encoder have the same output shape, which is (39,128). Their outputs are concatenated and sent to the decoder.

Decoder

The decoder comprises of two BiLSTMs layers and a dense layer. The decoder sends the combined output of the encoder to the first BiLSTMs with 256 units, and the output of the first BiLSTMs is fed to the second BiLSTMs with 512 units. The output of the second BiLSTMs is finally sent to a dense layer with a softmax activation function for word prediction. The model architecture of the encoder-decoder is shown in Fig. 4.

V. Experiments

In this section, a detailed discussion of the experimental details during training and validation of the proposed model architecture for Bangla image captioning is provided.

Experimental Setup

During training of the model, the hardware configuration comprised of a Ryzen 7 3700x CPU, 16GB DDR4 RAM, and Nvidia GTX 1070 8GB graphics card. We implemented our encoder-decoder model using the Tensorflow 2.6 deep learning framework within the Python 3.8 programming language. We deployed the trained Bangla image captioning model onto a Raspberry Pi 4 model B with 8GB RAM.

Parameter Setting

During training, $categorical\_crossentropy$ is used as loss function. Batch size is set to 485. Moreover, $RMSprop$ is selected as the optimizer. $RMSprop$ optimizer selects a distinct learning rate for each parameter during training, which significantly increases model performance. The weights of the model are updated following the Eq. 5 and Eq. 6 during training while using $RMSprop$ optimizer.

$$v_t = \beta_{t-1} + (1-\beta) * {g_t}^2 \tag{5}$$

$$W_{new} = W_{old} - \frac{n}{\sqrt{v_t + \epsilon}} * g_t \tag{6}$$

Here v_t is the average movement speed of gradient, g_t is the cost, β is the moving parameter and in the proposed model its value was 0.99. To calculate the new weights W_{new}, we subtract learning rate (η) times cost g_t which is divided by root over sum of v_t and a constant ϵ of very small value, from the old weights W_{old}.

Performance Metrics

To evaluate the performance of the proposed Bangla image captioning model, we calculate the BLEU score for the generated caption by the trained model. An individualised N-gram BLEU score is the assessment of matching grammes of a particular order, whereas cumulative BLEU scores relate to the computation of single n-gram scores for all orders from 1 to n. Consequently, the cumulative BLEU score is the most reliable metric for evaluating the real-world performance of a sentence generation algorithm. A cumulative BLEU score greater than 70 indicates that the sentence generated by the machine resembles a caption provided by a human. Calculation of cumulative BLEU score is given in Eq. 7, where c refers to length of predicted sentence and r refers to length of the original sentence and p stands for precision.

$$BLEU = min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{4} \frac{logP_n}{4} \tag{7}$$

VI. Result Analysis

In this section, the results found during the training and validation of tge proposed encoder-decoder model for Bangla image is presented. Moreover, we summarise the findings of our research work.

We trained our encoder-decoder model for 90 epochs. Fig. 5 and Fig. 6 exhibits the accuracy curve and the loss curve, respectively, and it is apparent from both curves that the model converges after 70 epochs. After 70 epochs of training, the proposed model attained an accuracy of 90% and a loss of less than 0.2.

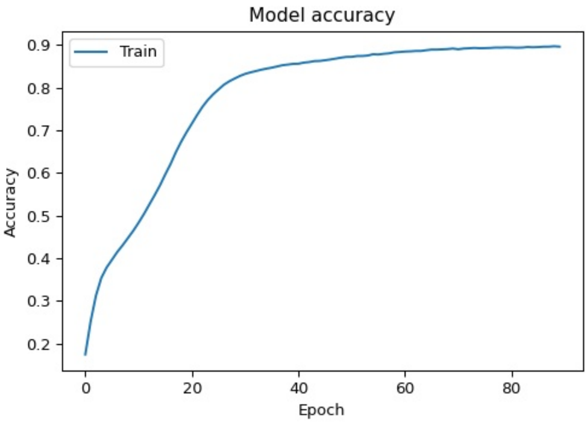


Fig. 5. Accuracy vs Epoch Curve.

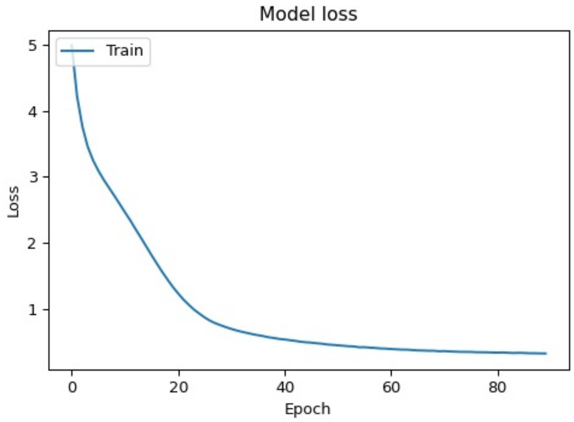


Fig. 6. Loss vs Epoch Curve.

We have implemented three popular pre-trained feature extractors namely VGG-16, Xception, and ResNet-50. Table II shows the BLEU score on the test dataset for the trained models using different pretrained feature extractors. It is found that ResNet-50 performs best as a feature extractor as compared to the VGG-16 or Xception model. The BLEU scores shown in Table II are cumulative BLEU scores.

From Fig. 7, it is evident that our encoder-decoder model with ResNet-50 pretrained model provides a significant improvement in BLEU score compared to the BLEU score stated in the previous research work on the same BanglaLekhaImageCaptions dataset. All of the machine-generated captions displayed in Table III are generated using test samples from the BanglaLekhaImageCaption dataset, which are unseen to the model during the training phase.

Table IV demonstrates that ResNet-50 outperforms the other two feature extractors when evaluated on new data from real-world scenarios. When VGG-16 and Xception are used as feature extractors, it is observed that the generated captions lack meaning and do not correspond to the meaning of the input image.

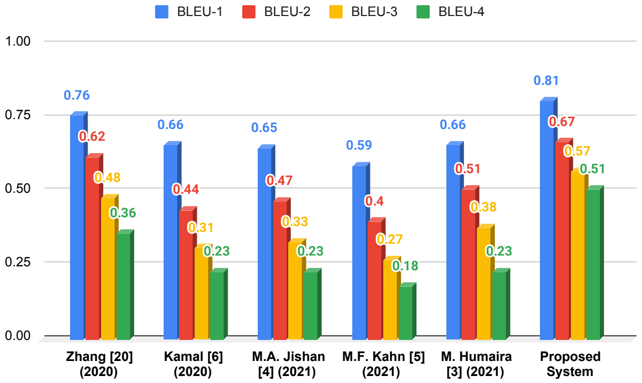
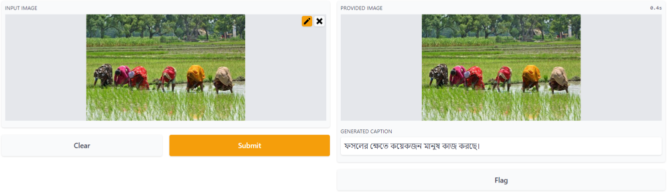


Fig. 7. Accuracy Comparison of Proposed System with Other Research.



ফসলের ক্ষেতে কয়েকজন মানুষ কাজ করছে।

Fig. 8. Web Integration.

For demonstration purposes, we have implemented the model in a web application. This model can also be used for information translation, to assist blind individuals in comprehending their surroundings by converting generated text to speech, and for many other purposes. Fig. 8 depicts an interface for Bangla image captioning within a primary web portal. When a web application receives an image as input, the image is sent to the back-end for image processing and text production. The resulting caption is then shown in the output textbox.

Finally, we have deployed our model on a Raspberry Pi 4 model B with 8GB of RAM. From Table V it is observed that the model only requires 122Mb of storage and the average inference time after testing 100 images was 400ms for ResNet-50 as feature extractor. In addition, the system consumes only 1000mA of current at full load. On the contrary, although flash occupation and energy consumption are similar for all three models, when ResNet-50 is used as a feature extractor, the model takes the least amount of average inference time to generate a caption.

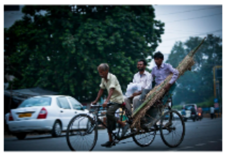
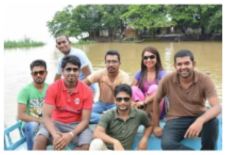
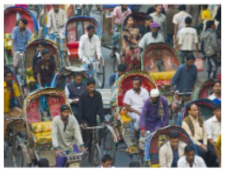
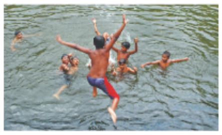
VII. Conclusion

In this research, we have successfully developed and deployed an encoder-decoder based deep learning model named "CapNet" that can generate syntactically and semantically correct and relevant Bangla captions from an input image and deployed the model into an embedded device. The model is trained on a public dataset named BanglaLekhaImageCaption

TABLE II. CUMULATIVE BLEU SCORES COMPARISON ON BANGLALEKHAIMAGECAPTIONS DATASET OF VARIOUS PRE-TRAINED MODEL USED IN ENCODERS OF THE PROPOSED MODEL

Experimental Models	BLEU-1 score	BLEU-2 score	BLEU-3 score	BLEU-4 score
VGG-16 with encoder-decoder	0.45	0.38	0.34	0.31
Xception with encoder-decoder	0.38	0.32	0.29	0.27
ResNet-50 with encoder-decoder	**0.81**	**0.67**	**0.58**	**0.51**

TABLE III. CAPTIONS GENERATED USING PROPOSED MODEL ON TEST-SET OF BANGLALEKHAIMAGECAPTION DATASET

Input Image	Generated Caption
	একজন পুরুষ রিক্সা চালিয়ে যাচ্ছে। (A man is driving a rickshaw.)
	কয়েকজন ছেলে ও একজন মেয়ে নৌকায় বসে আছে। (A few boys and a girl are sitting in the boat.)
	রাস্তায় অনেকগুলো রিক্সা চলছে যেগুলোতে অনেকগুলো মানুষ উঠে আছে। (There are many rickshaws on the road with many people on them.)
	কয়েকটি শিশু পানিতে লাফ দিচ্ছে। (A few children are jumping into the water.)

which has 9154 images with two captions per image. The model has attained the highest cumulative BLEU scores compared to all the previous works on Bangla image captioning that utilised this dataset. In addition, a comparison among three pre-trained image feature extraction models namely ResNet-50, VGG-16, and Xception is provided for the encoder network and it is found that ResNet-50 yields the best BLEU score for Bangla image captioning. Finally, we have deployed our model on a Raspberry Pi 4 model B with 8GB RAM and analysed the inference time and power consumption. In the future, we will enlarge the Bangla image captioning dataset by collecting and labelling more image data and training our model for achieving a higher BLEU score. We will deploy the trained Bangla image captioning model into an Android application so that visually impaired people can have a better understanding of their surroundings.

ACKNOWLEDGMENT

We appreciate the financial support provided by the Institute of Energy, Environment, Research, and Development (IEERD, UAP) and the University of Asia Pacific.

TABLE IV. CAPTIONS GENERATION COMPARISON USING THREE PRE-TRAINED MODELS AS FEATURE EXTRACTOR IN THE ENCODER

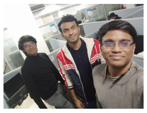
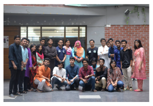
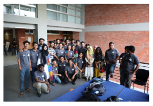
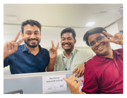
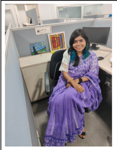
Input Image	ResNet50 incorporated model	Xception incorporated model	VGG-16 incorporated model
	তিনজন বন্ধু মিলে একসাথে দাঁড়িয়ে আছে। (Three friends are standing together.)	একজন পুরুষ সারা শরীরে দিয়েছে। (A man gave the whole body.)	২ জন ছেলে একটি লাইব্রেরীতে দাড়িয়ে ছবি তুলছে (2 boys are standing in a library taking pictures)
	অনেকগুলো নারী ও পুরুষ একসাথে আছে। (Many men and women are together.)	একটি মূর্তি আছে। (There is a statue.)	কিছু ছেলেমেয়ে একসাথে দাড়িয়ে এবং বসে আছেন একটি স্কুলের জায়গায়। (Some children are standing and sitting together in a school place.)
	কিছু মানুষ একটি ক্লাস এর ছাদে দাড়িয়ে এবং বসে আছেন ছবি তোলার জন্য। (Some people are standing and sitting on the roof of a classroom to take pictures.)	ঘরের সামনে কয়েকজন নারী ও কয়েকজন পুরুষ দাড়িয়ে আছে। (A few women and a few men are standing in front of the house.)	অনেকগুলো পুরুষ গোল হয়ে বসে আছে। (Many men are sitting in a circle.)
	পাশাপাশি তিনজন পুরুষ দাঁড়িয়ে আছে। (Three men stand side by side.)	একজন পুরুষ বসে ছবি তুলছে। কয়েকজন বালিকা দাঁড়িয়ে আছে (A man is sitting and taking pictures. Some girls are standing)	একজন বয়স্ক পুরুষ ও একজন নারী আছে। (There is an old man and a woman.)
	একজন নারী বসে আছে। (A woman is sitting.)	দুইজন পুরুষ একজন পুরুষকে এ পানিতে নেমে আছে (Two men are pushing a man into the water)	২ জন ছেলে এবং ১ জন হাতে একটি ব্যাগ নিয়ে বসে আছে একটি ফুলের বেঞ্চে একটি কপালে বই ধরে চেয়ে আছে । (2 boys and 1 with a bag in hand are sitting on a flower bench holding a book on their forehead.)

TABLE V. INFERENCE TIME, ENERGY CONSUMPTION AND FLASH OCCUPATION OF THE PROPOSED MODEL

Model	Flash occupancy (in Mb)	Inference time of proposed system (in ms)	Energy consumption (in mJ)
ResNet-50 with encoder-decoder	122	400	1000
VGG-16 with encoder-decoder	121	1100	1000
Xception with encoder-decoder	119	800	1000

REFERENCES

[1] S. Li, G. Kulkarni, T. Berg, A. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, 2011, pp. 220–228.

[2] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 684–699.

[3] M. Humaira, S. Paul, M. Jim, A. S. Ami, and F. M. Shah, "A hybridized deep learning method for bengali image captioning," *IJACSA*, vol. 12, no. 2, pp. 698–707, 2021.

[4] M. A. Jishan, K. R. Mahmud, A. K. Al Azad, M. R. Ahmmad, B. P.

Rashid, and M. S. Alam, “Bangla language textual image description by hybrid neural network model,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 2, pp. 757–767, 2021.

[5] M. F. Khan, S. Sadiq-Ur-Rahman, and M. S. 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*. Springer, 2021, pp. 217–229.

[6] A. H. Kamal, M. A. Jishan, and 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.

[7] M. A. H. Palash, M. Nasim, S. Saha, F. Afrin, R. Mallik, and S. Samiappan, “Bangla image caption generation through cnn-transformer based encoder-decoder network,” *arXiv preprint arXiv:2110.12442*, 2021.

[8] Y. Ming, N. Hu, C. Fan, F. Feng, J. Zhou, and H. Yu, “Visuals to text: A comprehensive review on automatic image captioning,” *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 8, pp. 1339–1365, 2022.

[9] G. Luo, L. Cheng, C. Jing, C. Zhao, and G. Song, “A thorough review of models, evaluation metrics, and datasets on image captioning,” *IET Image Processing*, vol. 16, no. 2, pp. 311–332, 2022.

[10] V. Ordonez, G. Kulkarni, and T. Berg, “Im2text: Describing images using 1 million captioned photographs,” *Advances in neural information processing systems*, vol. 24, 2011.

[11] R. Lebret, P. Pinheiro, and R. Collobert, “Phrase-based image captioning,” in *International Conference on Machine Learning*. PMLR, 2015, pp. 2085–2094.

[12] M. Rahman, N. Mohammed, N. Mansoor, and S. Momen, “Chittron: An automatic bangla image captioning system,” *Procedia Computer Science*, vol. 154, pp. 636–642, 2019.

[13] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, “Image captioning: Transforming objects into words,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[14] W. Lan, X. Li, and J. Dong, “Fluency-guided cross-lingual image captioning,” in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1549–1557.

[15] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, “Stair captions: Constructing a large-scale japanese image caption dataset,” *arXiv preprint arXiv:1705.00823*, 2017.

[16] V. Jindal, “Generating image captions in arabic using root-word based recurrent neural networks and deep neural networks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[17] S. R. Laskar, R. P. Singh, P. Pakray, and S. Bandyopadhyay, “English to hindi multi-modal neural machine translation and hindi image captioning,” in *Proceedings of the 6th Workshop on Asian Translation*, 2019, pp. 62–67.

[18] A. Jaffe, “Generating image descriptions using multilingual data,” in *Proceedings of the second conference on machine translation*, 2017, pp. 458–464.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[20] T. Deb, M. Z. A. Ali, S. Bhowmik, A. Firoze, S. S. Ahmed, M. A. Tahmeed, N. Rahman, and R. M. Rahman, “Oboyob: A sequential-semantic bengali image captioning engine,” *Journal of Intelligent & Fuzzy Systems*, vol. 37, no. 6, pp. 7427–7439, 2019.

[21] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.