Journal of King Saud University – Computer and Information Sciences 35 (2023) 257–269

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

An attention-based hybrid deep learning approach for bengali video captioning

Md. Shahir Zaoad, M.M. Rushadul Mannan, Angshu Bikash Mandol, Mostafizur Rahman, Md. Adnanul Islam, Md. Mahbubur Rahman *

Department of Computer Science and Engineering, Military Institute of Science and Technology, Dhaka 1216, Bangladesh

ARTICLE INFO

Article history:
Received 2 July 2022
Revised 6 November 2022
Accepted 25 November 2022
Available online 5 December 2022

Keywords:
Bengali video captioning
Convolutional neural network
Encoder-decoder model
Recurrent neural network
Attention-mechanism

ABSTRACT

Video captioning is an automated process of captioning a video by understanding the content within it. Although numerous studies have been performed on video captioning in English, the field of video captioning in Bengali remains nearly unexplored. Therefore, this research aims at generating Bengali captions that plausibly describe the gist of a specific video as well as identifying the best performing model for Bengali video captioning. To accomplish this, several sequence-to-sequence models – LSTM, BiLSTM, and GRU are implemented that takes the video frame features as input, extracted through different CNN models – VGG-19, Inceptionv3, and ResNet50v2, and provides a corresponding textual description as output. Moreover, the Attention mechanism is incorporated with these models as a first-ever attempt in Bengali video captioning. In this study, a novel Bengali video captioning dataset is constructed from Microsoft Research Video Description Corpus (MSVD) dataset (an English video captioning dataset) through utilizing a deep learning-based translator and manual post-editing efforts. Finally, the model's performance is evaluated in terms of popular performance evaluation metrics - BLEU, METEOR, and ROUGE. The proposed attention-based hybrid model outperforms the existing models in terms of these evaluation metrics, establishing a new benchmark for Bengali video captioning.
© 2022 The Author(s). Published by Elsevier B.V. on behalf of King Saud University. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Video captioning is a process of generating textual descriptions spontaneously from videos (Ji et al., 2022). In comparison with image captioning (Ali et al., 2022), the scenes in a video vary more frequently. A video also contains more information than a static image. Therefore, the task of video captioning poses more challenges than image captioning. The process of video captioning combines Natural Language Processing (NLP) and Computer Vision (CV) techniques. This process can be divided into two tasks, video feature extraction, and video description. The task of generating textual descriptions automatically from the videos aims to make a machine aware of the context in a given video by bridging the video content and its corresponding description with semantic consistency. With the development of sophisticated capturing devices and display techniques, the rapid growth of making video content can also be noticed for multifarious purposes, including news, emergencies, weather forecasts etc. Each and every passing moment, a significant number of new video contents are emerging on social media platforms namely Youtube and Facebook, which are crucial to be analyzed right away for the betterment of local and global communities. In order to confront this challenge, the most effective means is the automatic generation of captions for describing images and videos.

Most of the works of video captioning are focused on the English language. Even though Bengali is one of the most spoken languages (7th worldwide with 268 million speakers), only a limited number of work is present regarding Bengali language (Khairullah, 2019; Islam et al., 2022). The underlying reasons could be the scarcity of suitable Bengali datasets (i.e., extremely low-resource) (Mukta et al., 2021) or the complexity associated with Bengali video captioning itself. However, video captioning in Bengali can significantly help towards the advancement of various application domains, specially for the vast Bengali-speaking community. To address this gap, this study tests three different CNN

* Corresponding author.
E-mail addresses: adnanul@cse.mist.ac.bd (M.A. Islam), mahbub@cse.mist.ac.bd (M.M. Rahman).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2022.11.015
1319-1578/© 2022 The Author(s). Published by Elsevier B.V. on behalf of King Saud University.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

(VGG19, ResNet50v2 and Inception v3) models with the stated RNN models (GRU, LSTM and BiLSTM) to determine the best CNN-RNN combination for Bengali video captioning. While conventional RNNs are prone to vanishing gradient issues and have limited capability of dealing with information from many time steps, LSTM successfully overcomes these obstacles as it possess memory units to update hidden conditions and can handle long-term dependencies (Hochreiter and Schmidhuber, 1997). Similar to LSTM, GRU (Cho et al., 2014) also uses gates to control the flow of information. GRU has an overall simpler architecture, making it faster to train. On the other hand, Bidirectional LSTM (BiLSTM) (Bin et al., 2018) constructively increases the amount of information available to the network, which helps to improve the context accessible to the algorithm. Furthermore, as the first ever attempt in Bengali video captioning, this research also incorporates the attention mechanism (Luong et al., 2015; Xu et al., 2015; Yao et al., 2015; Yang et al., 2016) with the RNN architecture and makes the comparison to find out the most suitable combination with the attention mechanism for generating captions in Bengali. Another notable contribution is the identification of language effect that accounts for the performance difference among different RNN models in terms of Bengali video captioning. Finally, this research aims to develop a novel dataset[1] for Bengali video captioning from the MSVD dataset (Chen and Dolan, 2011), to identify the best performing model for Bengali video captioning from a combination of state-of-the-art models along with incorporating the attention mechanism with it. To recapitulate, the major contributions in this study are as follows:

- This study explores several state-of-the-art CNN and RNN based machine learning models to present the best CNN-RNN combination along with necessary configurations needed for Bengali video captioning.
- This study makes the first-ever attempt in incorporating the attention mechanism with those deep learning models for video captioning in Bengali, highlighting a comparative analysis of language effect of two different languages on the performance of those models.
- Finally, this research contributes in developing a novel dataset for Bengali video captioning, achieving benchmark results in Bengali video captioning.

2. Literature review

With the recent success of activity recognition of videos (Jaouedi et al., 2020), research in the video captioning field has exponentially increased throughout past decades. In past decades, LSTM-based various video captioning models were introduced as an effective replacement of conventional RNN where they successfully address the hurdles of conventional RNNs by incorporating memory units and being capable of handling long-term dependencies (Hochreiter and Schmidhuber, 1997). Moreover, Bidirectional LSTM (BiLSTM), an improvement of the same model, was introduced which effectively increases the amount of information available to the network by implementing LSTM architecture in both forward and backward directions (Islam et al., 2021). BiLSTM being a joint visual modeling approach is designed to encode video data while addressing both forward and backward dependencies by utilizing a forward LSTM pass along with a backward LSTM pass. The recent breakthrough in video captioning architecture came by introducing an attention mechanism into the encoder-decoder framework (Gao et al., 2017). With the attention mechanism, the salient features can be selected, and also the correlations between sentence semantics and visual content can be considered.

Although different frames attribute differently in feature learning, majority of the existing deep learning models assign equal weights irrespective of different visual and temporal features. This tremendously affects the feature distinction determination. In order to overcome this visual attention ignoring issue, Dai et al. (2020) proposed an end-to-end two-stream attention-based LSTM network that utilizes the visual attention mechanism, which selectively focuses on the useful features for the original input. A recent study by Ji et al. (2022), proposed an attention-based dual learning (ADL) approach that extracts the information from input videos along with the generated captions for video captioning, where a caption generation module and a video reconstruction module were implemented with the multi-head attention mechanism. Besides, Lin et al. (2022) proposed an end-to-end transformer-based video captioning model, which receives video frame patches as inputs and generates a natural language description. They show a comparison between sparsely sampled video frames and densely sampled video frames and concludes that in contrast to recent results with sparsely sampled video frames, video captioning can gain greatly from more densely sampled video frames (e.g., video question answering).

There are several studies in video captioning considering languages other than English, such as Chinese and Hindi. For example, Singh et al. (2021) suggested a hybrid attention mechanism for video captioning in Hindi by devising a soft temporal attention mechanism coupled with a semantic attention mechanism in order to make the system flexible in terms of focusing between the visual context vector and semantic input. Additionally, Wang et al. (2019) focused on preparing a large-scale, high-quality multilingual dataset for video-and-language research, containing over 41250 video segments, each video clip illustrating a single activity. Also, each video clip has 10 English descriptions and 10 Chinese descriptions. In VATEX Captioning Challenge 2020, Lin et al. (2020) demonstrated a video captioning model that combines a multi-modal feature fusion system integrated with feature attention. Their proposed model outperformed the official baseline on the English and the Chinese private test sets with a significant gap based on CIDEr metrics. To increase the encoder's representation capabilities, they retrieved multi-model features from the motion, appearance, semantic, and auditory domains. They utilized a two-layer GRU with an attention module as decoder.

Though extensive amount of work can be found regarding video captioning in different languages, video captioning in Bengali remains nearly unexplored. The first shortcoming in Bengali video captioning task is the lack of publicly available and semantically sound large datasets. A relevant work was done by Jishan et al. (2021) recently, which introduced a Bengali image captioning dataset, named as Bangla natural language image to text (BNLIT) dataset and proposed a hybrid encoder-decoder model consisting of a CNN using an encoder and a combination of BRNN (Bi-directional recurrent neural network) and LSTM language model for the sentence representation. The proposed model could detect images with multimodal and semantic complications. Palash et al. (2022) proposed a transformer-based architecture with an attention mechanism for Bengali image captioning. For feature extraction, it used a pre-trained ResNet-101 encoder model over the BanglaLekhaImageCaptions dataset. Besides, Shah et al. (2022) utilized three different Bengali datasets (Flicker8k-BN, BanglaLekha, and a Combined dataset) to generate Bengali caption using a visual attention-based encoder-decoder approach. Their comparison between the transformer-based model and other models clearly illustrates that the transformer-based model can uplift the original performance and improve the training speed by allowing parallelism. Although these works provide ideas about Bengali

[1] https://github.com/rushadmannan/BVC-with-attention-based-mechanism.git.

language-specific feature extraction and captioning techniques, they do not specifically deal with video captioning in Bengali.

To this end, the only work was done by Raj et al. (2021). This work trained and evaluated the model on the converted MSVD dataset which was translated using the help of google translator API. Here an encoder-decoder-based deep architecture was proposed that incorporated a combination of different CNNs with BiLSTM as the encoder and two layers of LSTM stacked on top of each other as the decoder. However, major limitations of their proposed model include the lack of a large-scale effective dataset and the absence of the implemented attention mechanism, which we address in this study as well as presenting a comparative analysis with his work. Finally, the background study clearly identifies the lack of fine tuned Bengali version of video captioning datasets and investigation of performance of various promising deep learning models in terms of Bengali video captioning. Therefore, this study aims at addressing these issues through extensive experimentation and evaluation.

3. Methodology

Bengali video captioning provides a machine with the capability to generate Bengali captions through training of neural network models. This process starts with the generation of vector representation from captions followed by feature extraction. Using these preprocessed inputs the encoder-decoder based model is trained to learn the mapping function between input videos and corresponding Bengali captions. Video captioning deals with sequential data where input comprises sequence of video frames while the output is textual description in other term sequences of words. Therefore in order to map these input video frames to the output captions primarily sequence-to-sequence (seq2seq) model is incorporated.

3.1. Model architecture

A seq2seq model, being an end-to-end model, comprises two portions: an encoder that encodes input into a fixed sized context vector and a decoder that generates output caption by utilizing the context vector. Consequently, seq2seq models are also addressed as encoder-decoder models. Where single RNN based sequence prediction is confined by sequence length and order, the seq2seq model delivers us from this hurdle, making it ideal for NLP. Fig. 1 illustrates the basic architecture of the seq2seq model for video captioning in Bengali. This research aims to distinguish the best performing model from various salient encoder-decoder based RNN architectures which are LSTM, BiLSTM and GRU.

3.1.1. Long short-term memory

Long Short-Term Memory is an Artificial Recurrent Neural Network architecture (Hochreiter and Schmidhuber, 1997) that comprises feedback connection along with the feed-forward neural networks. Owing to its characteristic architecture, LSTMs can handle the connection between recent past information and present tasks even if the gap grows, therefore, thus, resolves the long-term dependency issue. In LSTM structure, information flows through a memory mechanism known as cells which can selectively differentiate between information to be propagated and information to be forgotten. The cell state can carry information about sequential data processing such as speech, video, text, etc. A cell state consists of Input gate (x_t), Output gate (h_t) and Forget gate (f_t) as shown in Fig. 2. In a LSTM cell, input gate is utilized to quantify the significance of new information of a particular timestamp. A forget gate is the decisive component that is responsible for discarding (forgetting) the information from the previous timestamp. And, the motive of the output gate is the selection of dominant information from the current LSTM cell and flow that as the output. The mathematical equations (Yang et al., 2018) of the LSTM architecture for corresponding gates are shown in Eq. 1–6. LSTM is included in this research because of the property of remembering patterns selectively for a long time. Moreover, LSTM architecture is convenient for classifying, processing large time-series data and predicting the appropriate outcome.

$$i_t = \sigma\left(x_i U^i + h_{t-1} W^i\right) \tag{1}$$

$$f_t = \sigma\left(x_t U^f + h_{t-1} W^f\right) \tag{2}$$

$$o_t = \sigma(x_t U^o + h_{t-1} W^o) \tag{3}$$

$$\tilde{C}_t = tanh(x_t U^g + h_{t-1} W^g) \tag{4}$$

$$C_t = tanh(x_t U^g + h_{t-1} W^g) \tag{5}$$

$$h_t = tanh(C_t) * \ o_t \tag{6}$$

3.1.2. Bidirectional LSTM

Bidirectional LSTM is a sequence-to-sequence processing model incorporated with two LSTMs that can resolve issues of the basic LSTM model. BiLSTM looks specifically the same as its unidirectional counterpart. Memory units in LSTM based video captioning models can carry forward information from the previous time steps in one direction, making the model less robust as there is no backward connection between two consecutive memory cells. This architecture effectively improves the amount of information available, improving the context convenient to the algorithm. BiLSTM architecture (Bin et al., 2018) allocates the flow of information in both directions in a neural network; as a result, input information is preserved. Fig. 3. represents the basic architecture (Basaldella et al., 2018) of the Bi-LSTM incorporated with the Forward Hidden State ($\overrightarrow{h_t}$), Backward Hidden State ($\overleftarrow{h_t}$), Input(x_t) and Output(y_t).

The mathematical equations for Bi-LSTM are represented in Eq. 7–9. This research study integrates BiLSTM to explore the forward and backward temporal information in a sequence of spatio-temporal frames of a video to generate the corresponding description.

$$\overrightarrow{h_t} = tanh\left(W_{x\overrightarrow{h}} x_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}\right) \tag{7}$$

$$\overleftarrow{h_t} = tanh\left(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}\right) \tag{8}$$

$$y_t = W_{\overrightarrow{h}y} \overrightarrow{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y \tag{9}$$

3.1.3. Gated recurrent unit

Gated Recurrent Unit (GRU) is an advancement of basic Recurrent Neural Network incorporated with a gating mechanism (Cho et al., 2014). The flow of information in GRU is controlled through gates in an identical manner to LSTM. However, it has a simpler architecture with fewer parameters than LSTM and a faster training capacity. Fig. 4 depicts the basic architecture (Siam et al., 2017) of a single GRU unit inclusive of an update gate (z_t), reset Gate (r_t), current memory content (h_t) and final memory at current time step (h_t). The mathematical equations for GRU are represented in Eqs. 10,13.

$$z_t = \sigma(W_{zx} x_t + W_{zh} h_{t-1} + b_z) \tag{10}$$

$$r_t = \sigma(W_{rx} x_t + W_{rh} h_{t-1} + b_r) \tag{11}$$

$$\tilde{h}_t = tanh(W_{ox} x_t + W_{oh} r_t h_{t-1} + b_o) \tag{12}$$

$$h_t = z_t h_{t-1} + (1 - z_t)\tilde{h}_t \tag{13}$$

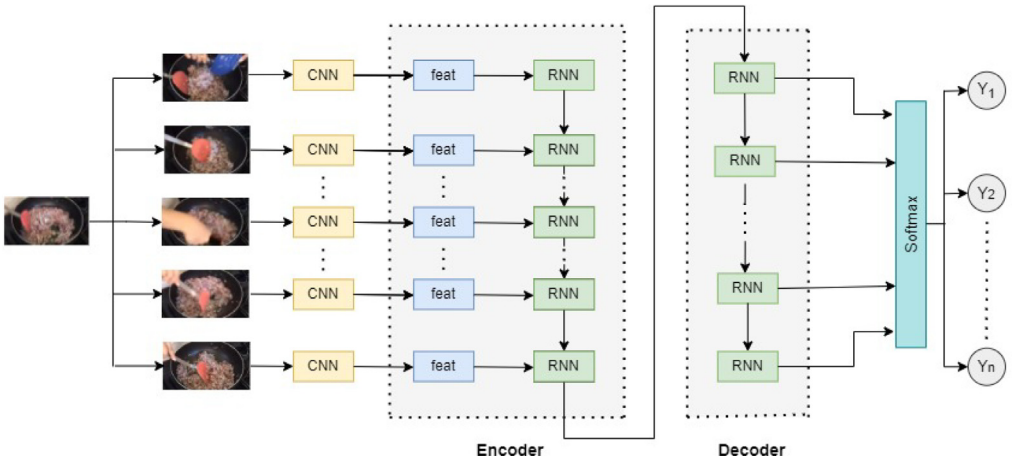


Fig. 1. Basic architecture of sequence-to-sequence model.

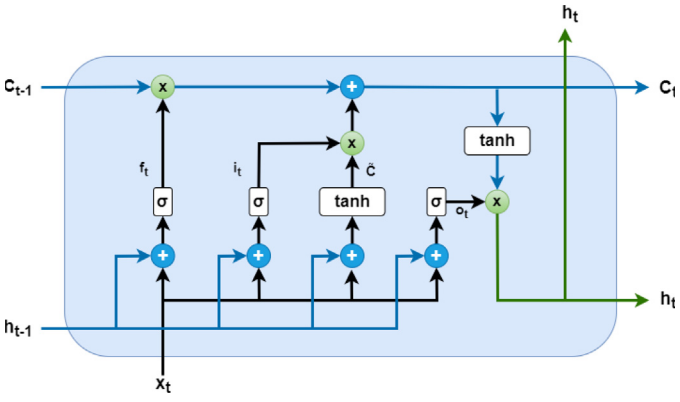


Fig. 2. Basic architecture of Long Short-Term Memory.

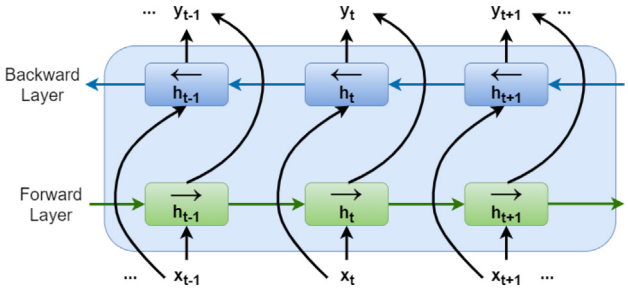


Fig. 3. Basic architecture of Bidirectional LSTM.

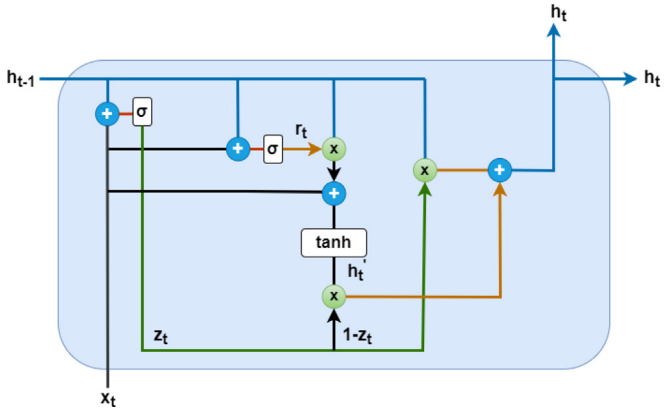


Fig. 4. Basic architecture of Gated Recurrent Unit.

This research incorporates GRU in the encoder-decoder portion of the generated sequence-to-sequence model, as GRUs have the tendency to exhibit better performance on smaller datasets with less frequent data points.

3.1.4. Attention mechanism

The shortcomings of encoder-decoder based seq2seq model is that it encodes the entire input sequence into one fixed length context vector. This is a matter of concern while dealing with longer input sequences, notably those that surpass ones in the training dataset. In order to resolve this limitation, the attention model was introduced where it develops a context vector specific to each decoder time step, unlike the general encoder-decoder model where the entire input sequence is mapped to a single context vector. Bahdanau Attention (Bahdanau et al., 2015) is used in this study, which is a general attention technique (attention mechanism depends on both input and output elements). Bahdanau Attention enables the flexible utilization for the decoder of the most relevant information of the input sequence. It does so by using all the encoder outputs to form a weighted combination where highest weights are assigned to the most relevant vectors. This attention mechanism is implemented within the three models discussed earlier (GRU, LSTM, BiLSTM) to distinguish the best performing one. Fig. 5 is the architecture of the implemented attention based seq2seq model.

In Fig. 5, up to the generation of encoder and decoder outputs, the process is similar to the general RNN model. In addition, the output from all the states of the encoder and the decoder are passed to the attention layer to generate attention output, followed by the concatenation of decoder outputs with attention output. Finally, the concatenated output is passed through the softmax layer to generate the prediction values.

3.2. Word embedding

Word embedding is required to represent a text in machine readable from where the words with similar meaning will have an identical representation. Such transformation is required in order to overcome the boundaries between human and machine comprehension. As a result, selected words from the training list are used to constitute a dictionary where each word is mapped using an integer vector, serving as the intermediary between human and computer.

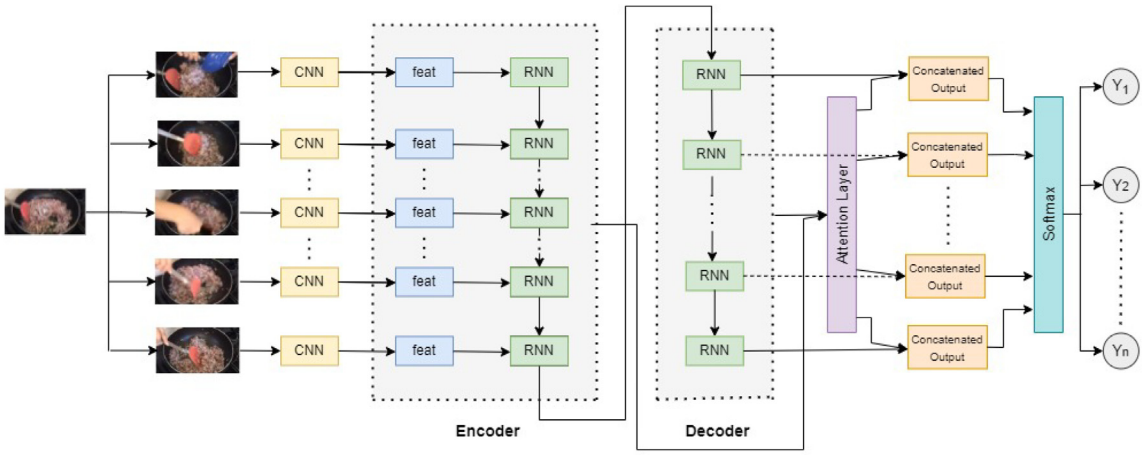


Fig. 5. Video Captioning using attention based RNN.

The tokenizer class of keras along with word2vec algorithm is used to generate a 1500 words long vocabulary where word2vec accounts for synonymous words. Each word from the training list is accompanied by a beginning of sentence token (<bos>) and an end of sentence token (<eos>) which is then checked for length validation. This research used captions with words between 6 to 10, thus discarding sentences with all other lengths. The selected sentences are used to extract the most frequent 1500 words and are assigned to the vocabulary which is a word index representation.

Finally, using this vocabulary, sentences are represented by corresponding integer index value followed by padding if necessary. In the training phase, the captions (words) are fed into the encoder LSTM as in this embedded form. Similarly using reverse mapping the vectorized values are used to output the predicted textual caption in Bengali.

3.3. Feature extraction

Like many other aspects that are intrinsic to human but difficult for machine is the apprehension of scenarios or in our case videos. In order to generate captions, first the machine has to understand the videos and distinguish among them. Therefore, pre-trained CNN models are incorporated in order to extract features from videos which helps a machine to understand video contents. This process involves splitting of each video into 80 frames where feature of each frame is extracted and represented in a mathematical notation (vector representation). Based on the CNN model used frames are scaled down to the size of 224 X 244 (VGG19 or ResNet) or 229 X 229 (Inception). Moreover, almost all the video in the dataset utilize RGB colors. Consequently, the input to the CNN model comprises a tensor of 224 X 224 X 3 or 229 X 229 X 3. After the completion of processing each video, the CNN model outputs a vector of (80 X P) where 80 corresponds to the frame number and "P" represents the number of features which varies upon the CNN used.

In this study, three different CNN (VGG19, ResNet50v2, Inception v3) models are tested with the stated RNN models (GRU, LSTM, BiLSTM) in order to distinguish the best CNN-RNN combination. These CNN models are notable for their speciality and contribution to feature extraction. The VGG19 model uses smallest kernel size of 3 X 3 in order to extract necessary information where 19 depicts the depth of the model. This model represents each video frame with a vector of 4096 values. On the other hand, Residual Network (ResNet) is able to introduce deep neural networks such as 50 or 100 layers, with skip connection technique, therefore these deep network is safe from vanishing gradient issue. And finally, in order to address images where the key object could have various representations, the Inception model combines filter of various sizes. Moreover, the output dimension for both theses latter CNN is a vector of size 2048; in other words, each of these CNNs represent a frame using 2048 values.

3.4. Encoder

Encoder is one of the key components of the RNN models. The encoder RNN is used to input the video frame features. In this research, RNN models are used as a many-to-many sequence model that implies, there is variation between input and output length. Therefore, the encoder encodes the input information into an intermediate representation which is then decoded by the decoder into an appropriate output sequence which may not be as the same length as the input sequence. And this intermediate representation is the last state of the encoder which is passed to the decoder as the initial state. Attributed to the number of video frames (80), the encoder architecture comprises 80 cells, one for each of the frames. In terms of encoder hidden layers, this research incorporated 512 of them. As a result, each encoder cell maps the video frame features to a vector of size 512 in a step.

The architecture of the encoder cell varies based on the RNN used. The main distinguishing component between LSTM and GRU is that LSTM cell comprises two internal states (cell state and hidden state) while GRU cell contains only hidden state. In terms of BiLSTM, the internal cell structure is identical to the LSTM but one additional LSTM layer is included in each encoder cell where through one LSTM layer the input flows in forward direction and backwards through the other. Furthermore, in case of the attention mechanism, the outputs from all the 80 encoder states are also considered rather than only the last step of the encoder.

3.5. Decoder

The other end of the RNN model is the decoder. This phase utilizes the encoder output vector (intermediate representation) which is the last stage of the encoder to predict the caption of a video. However, in case of learning, on top of the encoder output vector, the decoder cells intakes the vector representation of the reference captions in order to learn the mapping function. This process starts with, the first cell of the decoder taking <bos> (beginning of sentence) token to predict the first vector. Similarly, the second cell takes the output of the first decoder cell in order to predict the second vector, and so on. The process continues until the <eos> (end of sentence) token is reached.

The decoder model comprises 512 hidden states, i.e., all the vectors generated by the decoder are of size 512 and the general output dimension of the decoder is 10 X 512, where 10 represents the maximum length of the caption. This decoder output is passed through a dense layer with 1500 as the dimensionality of the output space in order to map the vector of size 512 to 1500 in order to match the vocabulary size. Softmax activation function is used in this layer as it predicts a multinomial probability distribution where it is required to predict multiple values in general.

3.6. Optimization function

Optimization refers to the phenomenon of minimizing the loss or in other terms maximizing the accuracy of the neural network model. In this study, Adam optimization is used (Kingma and Ba, 2014), where a distinct learning rate is maintained for individual network weight and is adapted according to the learning progression. Adam utilizes the pros of two other stochastic gradient descent extensions – Root Mean Square Propagation (RMSProp) and Adaptive Gradient Algorithm (AdaGrad). While a per-parameter learning rate is maintained by AdaGrad to improve performance on problems relating sparse gradients, RMSProp utilizes average of recent magnitudes of the gradients for the weight (e.g., how quickly it is changing) in order to update per-parameter learning rate. As a result, this algorithm performs well on noisy problems.

3.7. Loss function

The loss function is a measure used to evaluate the capability of a neural network to model the dataset. If the predictions are close enough to the actual value, then the loss function will output a lower number; otherwise, it will output a higher number. Different loss functions are concerned with different problems such as regression, binary and multiclass classification problems. Since video captioning is a multiclass classification problem, categorical cross-entropy is incorporated as a loss function (Islam et al., 2021). Categorical cross-entropy is also referred to as logarithmic loss where a prediction is penalized by comparing it with the actual class value and based on the distance from this actual value.

4. Experimental setup

Google colaboratory is used to conduct the majority of the experimentation regarding this work. This python development environment accommodates essential deep learning libraries namely NumPy, TensorFlow, and Keras, which are crucial for this study. On top of this, local machine is used, mostly for preparing the dataset along with feature extraction of the videos. Spyder, an Integrated Development Environment (IDE) is used as python development environment in the local machine.

4.1. Hyperparameter selection

One of the most significant determiner of a model's performance is hyperparameters which require external adjustment. Several key hyperparameters are considered in order to determine the best possible combination for any specific model. Table 1 represents the combinations of all the hyperparameters that are used to determine the best performing combination.

4.2. Dataset

A key concern regarding Bengali video captioning is the scarcity of proper Bengali dataset. Therefore, a notable contribution of this study is to present a syntactically and semantically consistent Bengali dataset for video captioning in Bengali. To this end, a widely used English dataset of video captioning, MSVD (Ji et al., 2022; Chen and Dolan, 2011), is utilized to present the corresponding Bengali captions. Primarily, a sophisticated deep learning-based translator[2] is used to automatically translate all 32,000 English captions of 1450 videos into Bengali captions. However, many translated captions are found erroneous with mostly two types of error. One type of the error is, some English words, especially slang words remained untranslated in Bengali. Another type of error is regarding Bengali prefixes, suffixes, and several special characters (Unicodes), causing disappearance of the associated valid letters and characters, as illustrated in Fig. 6. Moreover, some of the Bengali words contained or were replaced by English words, which is attributed to the inability of the translator to account for proper Bengali letters in such places. As a result, on top of the automatic translation, rigorous manual efforts were required to address these issues as well refining the quality of the machine translated Bengali captions. To do so, nine participants, having plausible proficiency in both Bengali and English languages, were involved in evaluation of the machine translated captions in order to compare the consistency based on human perceptions, ensuring the dataset's relevance to the natural Bengali language. Finally, this iteratively refined dataset is used to carry out the training and evaluation of the proposed video captioning models. In terms of training and testing, 1450 and 100 video snippets are used respectively, with an 85% split ratio for training and validation purposes.

Table 1
Combination of key hyperparameters.

Configuration	Epochs	Learning Rate	ReducedLROnPlateau
Configuration-1	100	0.0003	0.1
Configuration-2	100	0.0003	0.01
Configuration-3	50	0.00003	0.1

4.3. Training

The notion behind neural network mechanism is similar to that of human learning as first the machine is trained to apprehend various concepts which are then utilized to derive answers for unknown or new scenarios. In this phase of learning, under-the-hood what is really happening is a mapping between input and output which is in case of this proposed work, learning the relation between a video and corresponding caption. Owing to the fact that this research aims to determine the best performing combination (CNN and RNN), therefore, different salient deep learning models have been trained on the modified dataset to generate corresponding accuracy. Three selected CNN models (VGG19, ResNet50v2 and Inception v3) are applied along with RNN models (LSTM, GRU and BiLSTM) and further with attention mechanism. For training, an essential concern is regarding epochs, which is, a large number of them can cause overfitting, whereas too few can lead to underfitting of the model. In order to overcome this hurdle, early stopping (a callback approach) is incorporated which allows to define an arbitrarily large number of epochs given that it will automatically halt the training process with regards to a standstill situation regarding performance improvement.

For each combination of CNN and RNN, the best possible hyperparameter combination in terms of training accuracy is selected (see Table 2). These selected models are used for performance evaluation later.

[2] https://deep-translator.readthedocs.io/en/latest/.

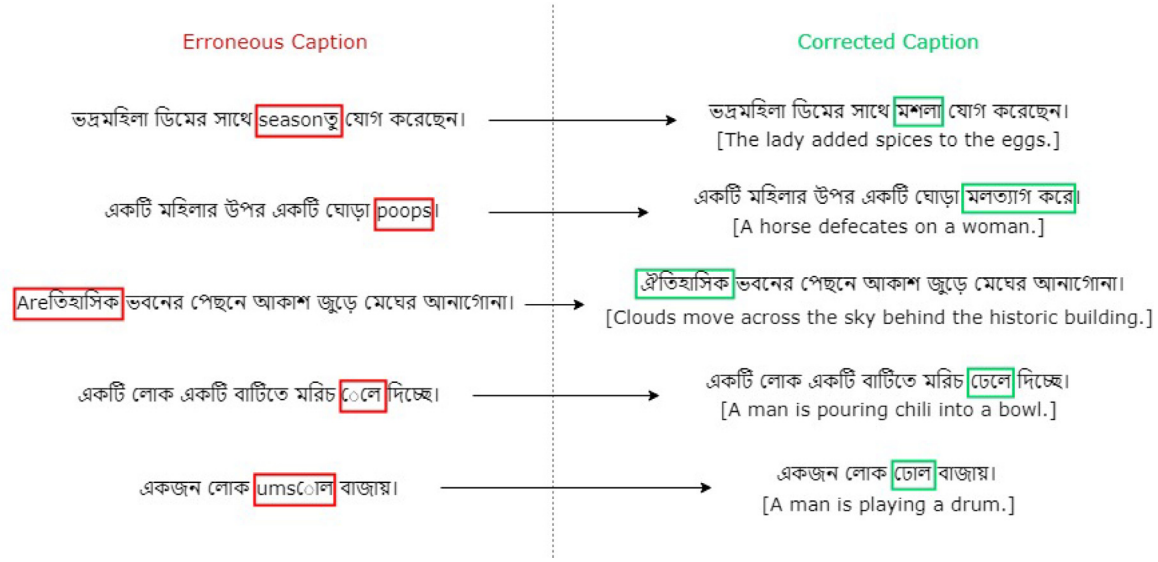


Fig. 6. Automatically generated erroneous captions and corresponding corrected ones. Corresponding English translations are provided as well (in square brackets) considering readers' convenience.

5. Result and evaluation

This section exhibits the outcome of the proposed models over the testing dataset and the selection of the best alternative model in terms of different evaluation metrics. Comprehensive comparisons among proposed models and existing Bengali video captioning models are also endorsed in this section. When evaluating video captioning models, some consistent evaluation protocols are used. Specifically, three popular metrics, widely used for measuring machine-generated text quality, are used in the evaluation of the proposed model: BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and ROUGE (Lin, 2004).

5.1. Quantitative analysis

Quantitative evaluation of this research is conducted on the testing dataset containing 100 video snippets from a minimum duration of 6 s to a maximum duration of 74 s. The evaluation is carried out for each combination of CNN and RNN models with the best hyperparameter combination. For all these models, incorporated with attention-based mechanism, captions have been generated using two different search algorithms. In order to generate captions in Bengali, the most likely output sequences have to be decoded by searching all possible output sequences. The proposed methodology incorporates beam search and greedy search techniques in this regard.

5.1.1. Performance of greedy search

The greedy search algorithm uses the local optimality to select a word for each stage to generate the caption. The simple approximation is to use the most likely word to decode the output sequence which contributes in terms of the speed of the decoder. By picking the single highest probability word one by one, greedy search has decoded a sentence in less than 5-7 s. However, the quality of the final output sequence may differ from the optimal sentence because this algorithm prefers multiple common words to rare words. Table 3 represents the performance scores using greedy search for each CNN and RNN combination (with and without attention mechanism) in terms of BLEU, METEOR and ROUGE evaluation metrics.

5.1.2. Performance of beam search

The beam search algorithm, a heuristic search method, selects K possible alternatives based on the conditional probability at each location for decoding an input sequence. The experimental setup for this research has utilized beam width of value 3 (k = 3), signifying that the search technique will consider three words to choose the most appropriate one. The search procedure halts for each sequence either by reaching the maximum length of 10 words or by reaching the end of the sequence token or the threshold likelihood. Length normalization helps the search method to generate better output sentences. On average, beam search has generated video captions in less than 45–70 s. BLEU, METEOR and ROUGE scores are shown in Table 4 (with and without attention mechanism) for the beam search method.

5.2. Qualitative analysis

This section presents the sample output of the proposed models having the best performances with and without the incorporation of the attention mechanism (Fig. 7). In Figs. 7(I), (II), and (III), the predicted captions convey the equivalent meaning of the actions performed in the video snippets and also very similar to the reference captions. The frames extracted from the videos have been able to perceive the correct information regarding the videos and the seq2seq model performed effectively to generate the captions.

The process of qualitative analysis has been performed on a subset of total 35 video clips and a set of reference captions for each clip. The validation of this analysis has been accomplished using the evaluation metrics. Considering the subset of videos, Fig. 8 illustrates a comparative analysis between greedy and beam search approaches in terms of different performance evaluation metrics.

5.3. Performance comparison

In this section, a comparison of the proposed models and the existing models regarding video and image captioning in Bengali is represented. In the field of video captioning, the Bengali language has seen too little work; however, several studies have been conducted on Bengali image captioning. Videos can relate to images, as it is a continuous sequence of images. Therefore, the

Table 2
Comparison of various CNN and attention based RNN combinations in terms of training accuracy and loss.

CNN	Attention + RNN	Hyperparameter Configuration	Accuracy	Val. Accuracy	Loss	Val. Loss
VGG19	LSTM	Configuration-2	0.7701	0.7059	0.7917	1.3311
ResNet50v2		Configuration-1	0.7591	0.6952	1.0352	1.4753
Inceptionv3		Configuration-2	0.7479	0.7009	1.0637	1.4216
VGG19	BiLSTM	Configuration-1	0.7951	0.7064	0.6739	1.3583
ResNet50v2		Configuration-2	0.7814	0.7066	0.8825	1.3970
Inceptionv3		Configuration-1	0.8066	0.7171	0.7078	1.3483
VGG19	GRU	Configuration-1	0.8005	0.7192	0.6685	1.2633
ResNet50v2		Configuration-2	0.7315	0.6912	1.1860	1.5121
Inceptionv3		Configuration-1	0.7691	0.7142	0.9628	1.3668

Table 3
Performance comparison of Greedy search for various CNN and RNN combinations (Bengali).

CNN	RNN	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE
VGG19	LSTM	**0.685**	**0.568**	**0.431**	**0.387**	**0.337**	**0.423**
ResNet50v2		0.363	0.339	0.287	0.240	0.167	0.334
Inceptionv3		0.408	0.321	0.300	0.285	0.197	0.363
VGG19	BiLSTM	0.337	0.302	0.253	0.185	0.128	0.356
ResNet50v2		0.500	0.408	0.340	0.320	0.220	0.38
Inceptionv3		0.544	0.433	0.387	0.309	0.215	0.356
VGG19	GRU	0.699	0.614	0.524	0.341	0.314	0.394
ResNet50v2		0.487	0.432	0.336	0.284	0.197	0.330
Inceptionv3		0.418	0.375	0.305	0.244	0.169	0.370
VGG19	LSTM + Attention	0.629	0.572	0.488	0.407	0.317	0.467
ResNet50v2		0.418	0.377	0.313	0.299	0.207	0.366
Inceptionv3		0.537	0.482	0.391	0.362	0.251	0.356
VGG19	BiLSTM + Attention	0.295	0.260	0.207	0.158	0.109	0.327
ResNet50v2		0.513	0.465	0.397	0.392	0.272	0.359
Inceptionv3		0.576	0.522	0.456	0.422	0.293	0.369
VGG19	GRU + Attention	**0.681**	**0.585**	**0.492**	**0.422**	**0.337**	**0.489**
ResNet50v2		0.502	0.460	0.387	0.309	0.214	0.350
Inceptionv3		0.571	0.522	0.448	0.430	0.298	0.366

The bold values represent the best performing combination of different CNN and RNN models.

Table 4
Performance comparison of Beam search for various CNN and RNN combinations (Bengali).

CNN	RNN	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE
VGG19	LSTM	**0.647**	**0.587**	**0.453**	**0.392**	**0.332**	**0.435**
ResNet50v2		0.378	0.337	0.268	0.265	0.176	0.317
Inceptionv3		0.570	0.509	0.410	0.339	0.226	0.357
VGG19	BiLSTM	0.309	0.284	0.233	0.197	0.136	0.336
ResNet50v2		0.418	0.392	0.363	0.307	0.224	0.351
Inceptionv3		0.493	0.438	0.408	0.329	0.252	0.336
VGG19	GRU	0.622	0.561	0.475	0.349	0.229	0.399
ResNet50v2		0.588	0.514	0.403	0.301	0.201	0.346
Inceptionv3		0.420	0.374	0.316	0.306	0.204	0.390
VGG19	LSTM + Attention	**0.711**	**0.580**	**0.507**	**0.434**	**0.391**	**0.472**
ResNet50v2		0.568	0.500	0.416	0.392	0.261	0.325
Inceptionv3		0.659	0.592	0.487	0.398	0.298	0.336
VGG19	BiLSTM + Attention	0.326	0.289	0.233	0.174	0.116	0.311
ResNet50v2		0.600	0.544	0.407	0.403	0.228	0.357
Inceptionv3		0.674	0.604	0.512	0.410	0.284	0.438
VGG19	GRU + Attention	0.693	0.608	0.501	0.428	0.360	0.470
ResNet50v2		0.626	0.564	0.455	0.330	0.220	0.330
Inceptionv3		0.630	0.548	0.435	0.352	0.281	0.420

The bold values represent the best performing combination of different CNN and RNN models.

performances of the proposed models are also compared with the studies accomplished for image captioning in Bengali. Table 5 exhibits the performance of different models with the proposed ones in terms of different evaluation metrics.

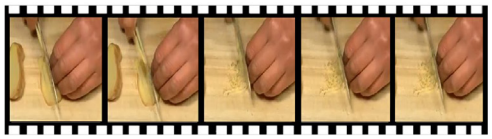
6. Exploring language effect in video captioning task

6.1. Performance of beam and greedy search over English captions

Similar to Bengali captions, this section exhibits the performance results generated from (equivalent) experiments for English captions. It is clear that attention based BiLSTM performs the best in terms of three models. The reason behind BiLSTM performing substantially better is attributed to its ability to propagate information in both forward and backward direction. Therefore, it can plausibly address the forward and backward dependencies in English sentences and thus, generating more accurate captions; presented in Tables 6 and 7.

6.2. Comparison between high-resource and low-resource contexts

One of the noteworthy findings considering a plausibly high-resource language context, English captions, is that opposed to the scenario for the low-resource context (Bengali captions), BiLSTM



Reference Caption:

1. একজন ব্যক্তি আলু টুকরো করছেন।(A man is chopping potatoes.)
2. কেউ ছুরি দিয়ে খুব ছোট টুকরো করে আদা টুকরো টুকরো করছে। (Someone is chopping ginger into very small pieces with a knife.)

Predicted Caption:

1. Without Attention: একজন মহিলা কিছু খাবার কাটছেন। (A woman is chopping off some foods.)
2. With Attention: একজন মহিলা আদা কাটছেন। (A woman is chopping ginger.)

(a)

Reference Caption:

1. একজন লোক ফোনে কথা বলছে। (A man is talking on the phone.)
2. একজন লোক ফোনে কথা বলছে এবং পায়চাড়ি করছে। (A man is talking on the phone and walking.)

Predicted Caption:

1. Without Attention: একজন লোক কথা বলছে। (A man is talking.)
2. With Attention: একজন লোক তার কথা বলছে। (A man is talking about himself.)

(b)

(I)

Reference Caption:

1. একজন মানুষ প্যানে পানির মধ্যে চাল ঢেলে নাড়ছে। (A man is stirring rice in a pan.)
2. একজন ব্যক্তি পাত্রের মধ্যে পানি এবং চাল ঢেলে দিচ্ছেন। (A man pours water and rice into a pot.)

Predicted Caption:

1. Beam Search: একজন ব্যক্তি একটি প্যানে চাল নাড়ছে। (A man is stirring rice in a pan.)
2. Greedy Search: একজন ব্যক্তি একটি প্যানে চাল নাড়াচ্ছেন। (A man is stirring rice in a pan.)

(c)

Reference Caption:

1. একজন মহিলা প্যানের মধ্যে বিভিন্ন উপাদান মিশ্রিত করে নেড়ে রান্না করছে। (A woman is cooking by mixing different ingredients in a pan.)
2. একজন ব্যক্তি ফ্রাইং প্যানে খাবার রান্না করছেন। (A man is cooking food in a frying pan.)

Predicted Caption:

1. Without Attention: একজন ব্যক্তি একটি প্যানে কিছু যোগ করছেন। (A person is adding something in the pan.)
2. With Attention: একজন মহিলা একটি প্যানে কিছু মাংস নাড়াচ্ছেন। (A woman stirring some meat in a pan.)

(d)

(II)

Reference Caption:

1. একটি ছেলে পিয়ানো বাজায়। (A boy plays the piano.)
2. একটি ছেলে একটি কী-বোর্ড খেলছে। (A boy is playing a keyboard.)

Predicted Caption:

1. Without Attention: একজন লোক পিয়ানো বাজায়। (A man plays the piano.)
2. With Attention: একজন মানুষ পিয়ানো বাজায়। (A man plays the piano.)

(e)

Reference Caption:

1. একজন মহিলা পেঁয়াজ কাটছেন। (A woman is chopping onions.)
2. একটি মহিলা একটি পেঁয়াজ কুচি করে। (A woman chops an onion.)

Predicted Caption:

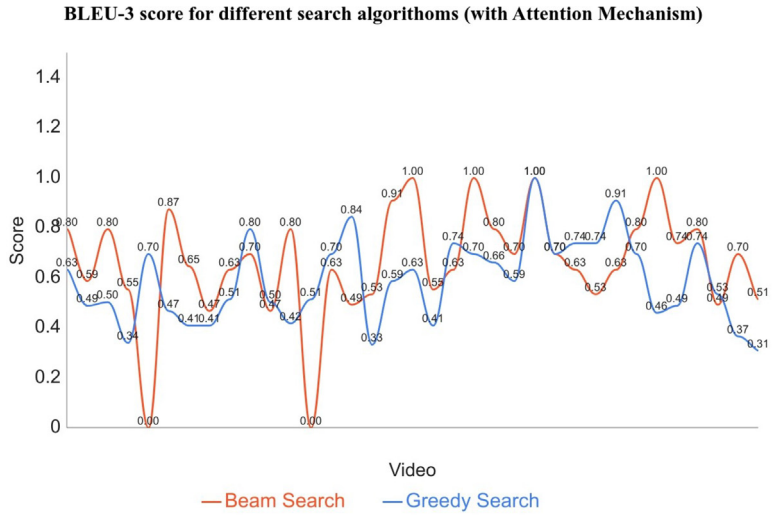
1. Without Attention: একজন মহিলা একটি ছোট টুকরো টুকরো করছেন। (A woman is doing a small piece.)
2. With Atttention: একজন মহিলা পেঁয়াজ কাটছেন। (A woman is chopping onions.)

(f)

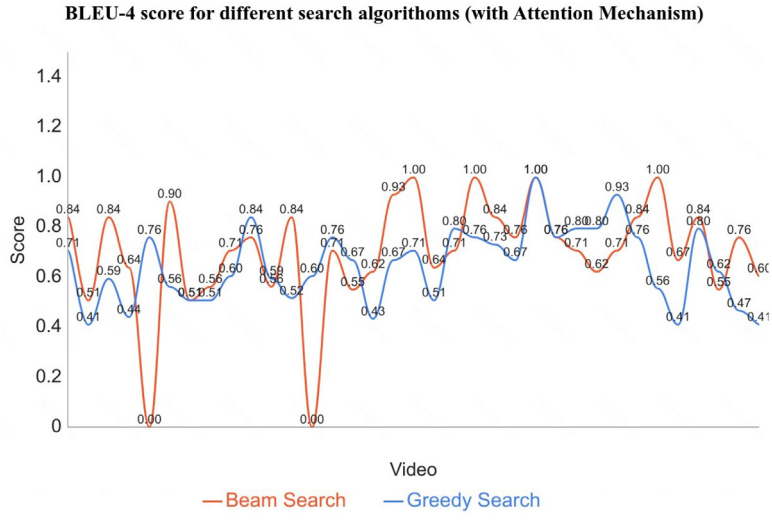
(III)

Fig. 7. Example of generated captions using the proposed models.

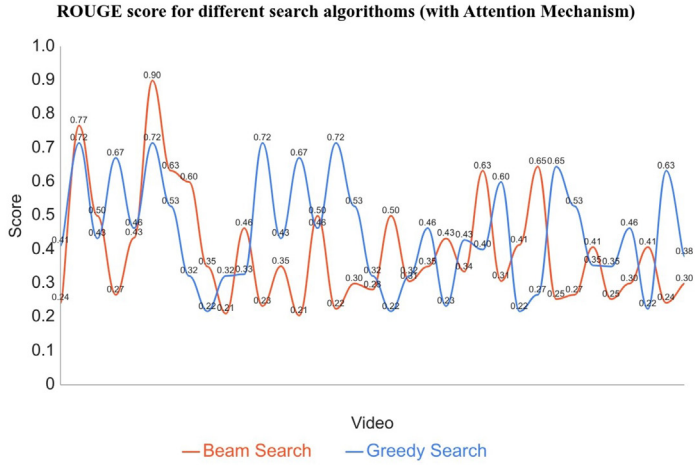
performs better than other models. The primary reason behind this phenomenon is, while English sentences have both forward and backward dependencies, Bengali sentences only have forward dependencies with hardly having any backward dependencies. As a result, BiLSTM, being capable of taking consideration of both forward and backward dependencies through its ability of forward and backward propagation, generates more semantically plausible English captions, ultimately attributing to higher evaluation result, as presented in Fig. 9(I) (considering Greedy search technique) and 9(II) (considering Beam search technique). On contrary, while working with Bengali captions, BiLSTM misinterprets some sentence logic and semantics because of its backward propagation; consequently, producing sentences, not properly reflecting the content of the video, thus resulting in poor evaluation result. However, Bengali significantly outperforms English in terms of LSTM and GRU, as only having forward dependencies in the sentences; therefore, these models can efficiently represent the relation present in Bengali sentences, as illustrated in Fig. 10(I) and (III).



(I)



(II)



(III)

Fig. 8. Qualitative assessment validation in terms of evaluation metrics.

Table 5
Performance comparison of proposed models with the existing models.

Relevant Works	Input Sequence	BLEU3	BLEU4	METEOR	ROUGE
Proposed Model	video	**0.453**	**0.392**	**0.332**	**0.435**
Proposed Model + Attention	video	**0.507**	**0.434**	**0.391**	**0.472**
Deep learning based approach (Raj et al., 2021)	video	0.432	0.326	-	0.573
BNLIT (Jishan et al., 2021)	image	0.324	0.228	-	-
BanglaLekha Image captioning (Kamal et al., 2020)	image	0.317	0.238	-	0.573
Bornon (Shah et al., 2022)	image	0.510	0.440	0.360	-
Flicker8k-BN (Humaira et al., 2021)	image	0.330	0.220	0.460	0.540

The bold values represent the best performing combination of different CNN and RNN models.

Table 6
Performance comparison of Greedy search for various CNN and RNN combinations (English).

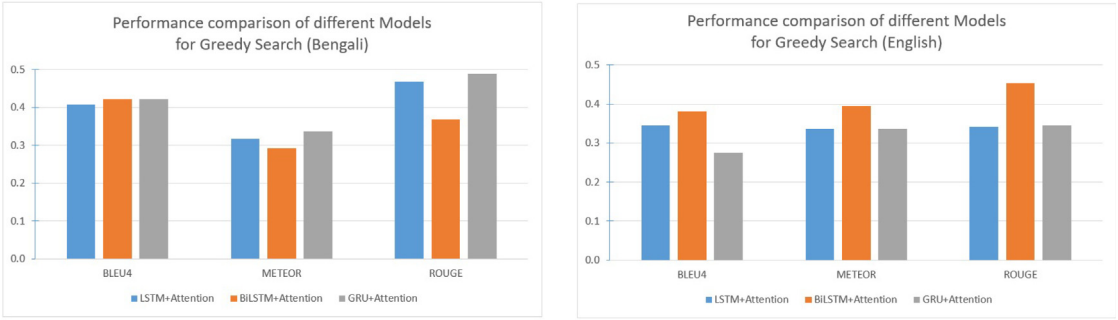
CNN	RNN	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE
VGG19	LSTM + Attention	0.697	0.569	0.423	0.345	0.337	0.342
ResNet50v2		0.671	0.522	0.367	0.311	0.271	0.310
Inceptionv3		0.684	0.492	0.326	0.308	0.389	0.307
VGG19	BiLSTM + Attention	**0.693**	**0.586**	**0.498**	**0.381**	**0.395**	**0.454**
ResNet50v2		0.675	0.490	0.387	0.322	0.372	0.383
Inceptionv3		0.652	0.473	0.375	0.363	0.337	0.372
VGG19	GRU + Attention	0.702	0.405	0.348	0.275	0.337	0.345
ResNet50v2		0.596	0.379	0.314	0.262	0.310	0.294
Inceptionv3		0.632	0.429	0.333	0.243	0.265	0.368

The bold values represent the best performing combination of different CNN and RNN models.

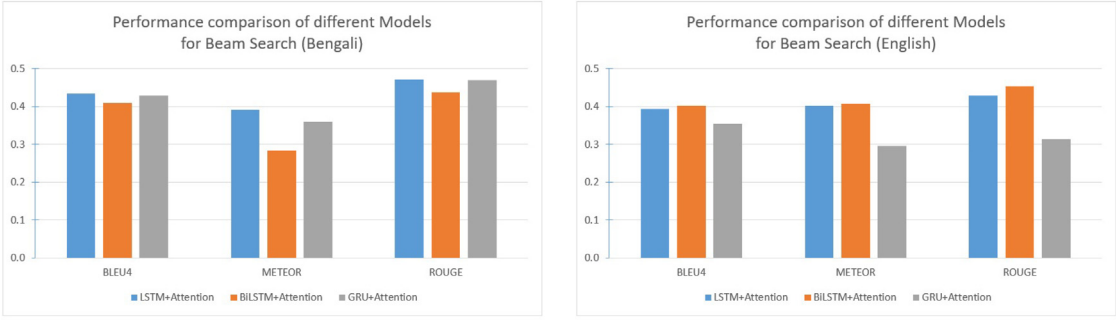
Table 7
Performance comparison of Beam search for various CNN and RNN combinations (English).

CNN	RNN	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE
VGG19	LSTM + Attention	0.725	0.599	0.521	0.393	0.403	0.428
ResNet50v2		0.651	0.467	0.408	0.329	0.311	0.309
Inceptionv3		0.685	0.592	0.469	0.352	0.317	0.369
VGG19	BiLSTM + Attention	**0.734**	**0.608**	**0.517**	**0.402**	**0.407**	**0.453**
ResNet50v2		0.679	0.563	0.469	0.336	0.363	0.427
Inceptionv3		0.698	0.587	0.483	0.389	0.380	0.426
VGG19	GRU + Attention	0.710	0.522	0.468	0.349	0.327	0.359
ResNet50v2		0.657	0.547	0.479	0.354	0.296	0.313
Inceptionv3		0.628	0.482	0.392	0.315	0.316	0.311

The bold values represent the best performing combination of different CNN and RNN models.



(I)



(II)

Fig. 9. Performance comparison between high-resource and low-resource context in terms of various Attention based RNN models.

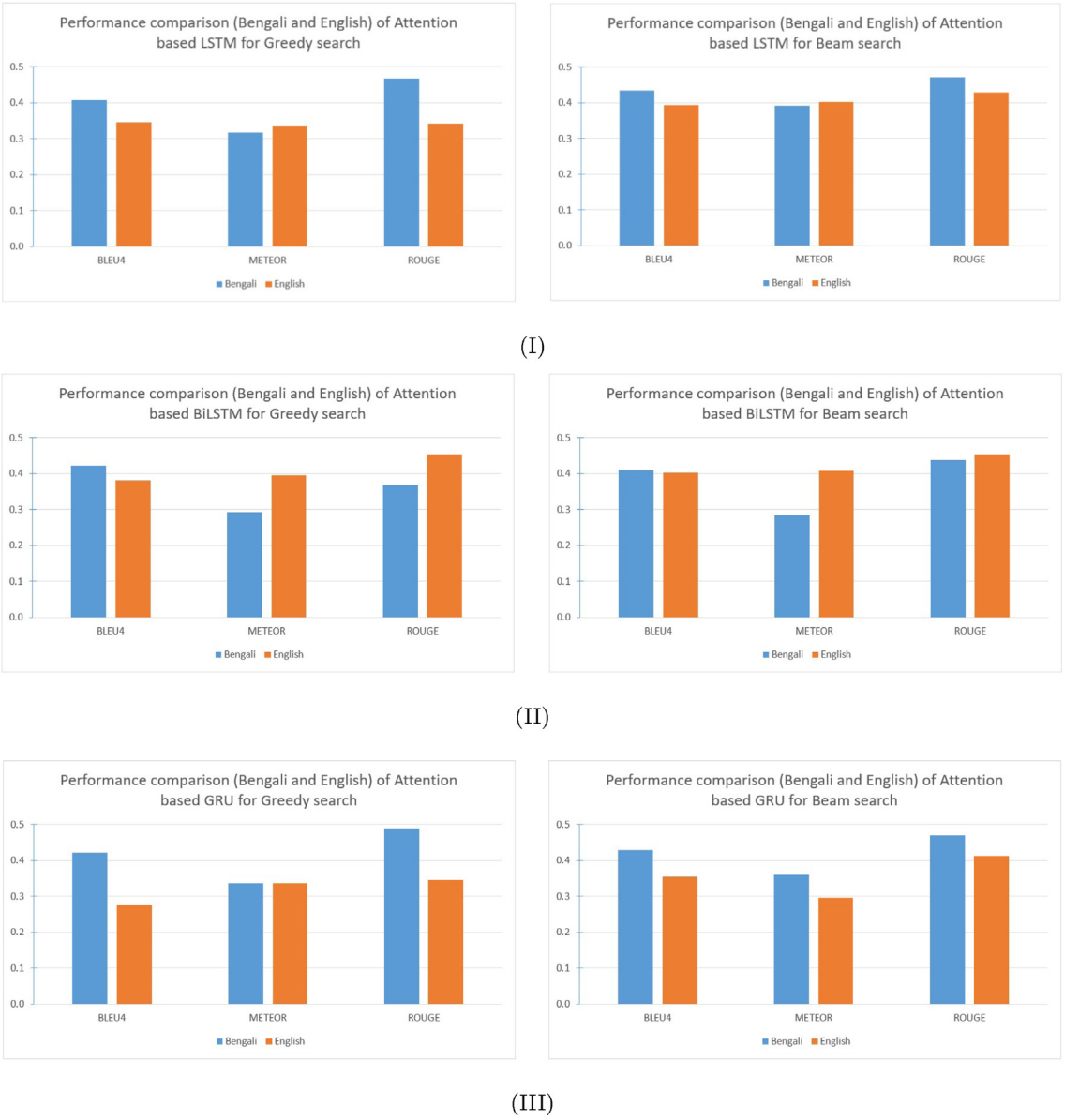


Fig. 10. Performance comparison between Beam and Greedy Search in terms of high-resource and low-resource context.

7. Conclusion

In this research, a rigorous experimentation was conducted among various salient neural network architectures and thus, the best performing model in terms of Bengali video captioning was identified. Another notable contribution of this research is the development of a syntactically and semantically consistent dataset from the MSVD that can be utilized for future research in this field. The selected state-of-the-art models are trained on the dataset in order to distinguish the model with maximum accuracy. Furthermore, the attention mechanism is introduced to achieve a benchmark performance for Bengali video captioning. LSTM along with VGG19 stands as the best performing model in terms of generic RNN architecture, while the soundness of the attention based GRU accompanied by VGG19 made it possible to generate captions that are more natural, thus making the overall best performing model. Finally, a comparative analysis is performed on the model's performance for two different search techniques in terms of three popular evaluation metrics - BLEU, METEOR and ROUGE, in order to carry out a thorough and versatile performance evaluation.

Although this research established a sound Bengali dataset for video captioning, further endeavour is required to establish a significantly prolonged size of the corpus that will help towards training a machine more rigorously. A notable concern regarding Bengali video captioning is the hurdle that is introduced because of specific language features of Bengali, which has been plausibly addressed in this study. However, more focused work is required to mitigate this effect, specifically for syntactically and semantically rich low-resource languages. Moreover, development of models for longer video clips, generating longer captions as well as detecting more than a single action, is another potential future work of this study.

Declarations

Funding. The authors did not receive support from any organization for the submitted work. No funding was received to assist with the preparation of this manuscript. No funding was received for conducting this study. No funds, grants, or other support was received.

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.

References

Ali, U., Lee, I.H., Mahmood, M.T., 2022. Robust regularization for single image dehazing. Journal of King Saud University - Computer and Information Sciences 34, 7168–7173.

Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. In: 2015, 3rd International Conference on Learning Representations, ICLR 2015.

Banerjee, S., Lavie, A., 2005. 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, pp. 65–72.

Basaldella, M., Antolli, E., Serra, G., Tasso, C., 2018. Bidirectional lstm recurrent neural network for keyphrase extraction. In: Italian Research Conference on Digital Libraries, Springer, pp. 180–187.

Bin, Y., Yang, Y., Shen, F., Xie, N., Shen, H.T., Li, X., 2018. Describing video with attention-based bidirectional lstm. IEEE transactions on cybernetics 49 (7), 2631–2641.

Chen, D., Dolan, W.B., 2011. Collecting highly parallel data for paraphrase evaluation, in. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, pp. 190–200.

Cho, K., van Merriënboer, B., Bahdanau, D., Bengio, Y., 2014. On the properties of neural machine translation: Encoder–decoder approaches. In: Proceedings of SSST 2014 - 8th Workshop on Syntax, Semantics and Structure in Statistical Translation, Association for Computational Linguistics (ACL), pp. 103–111.

Dai, C., Liu, X., Lai, J., 2020. Human action recognition using two-stream attention based lstm networks. Applied soft computing 86, 105820.

Gao, L., Guo, Z., Zhang, H., Xu, X., Shen, H.T., 2017. Video captioning with attention-based lstm and semantic consistency. IEEE Transactions on Multimedia 19 (9), 2045–2055.

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

Humaira, M., Paul, S., Jim, M.A.R.K., Ami, A.S., Shah, F.M., 2021. A hybridized deep learning method for bengali image captioning. Int. J. Adv. Comput. Sci. Appl. 12, 698.

Islam, M.A., Anik, M.S.H., Islam, A.A.A., 2021. Towards achieving a delicate blending between rule-based translator and neural machine translator. Neural Computing and Applications 33 (18), 12141–12167.

Islam, M.A., Anik, M.S.H., Islam, A.A.A., 2022. An enhanced rbmt: When rbmt outperforms modern data-driven translators. IETE Technical Review, 1–12.

Jaouedi, N., Boujnah, N., Bouhlel, M.S., 2020. A new hybrid deep learning model for human action recognition. Journal of King Saud University - Computer and Information Sciences 32 (4), 447–453.

Ji, W., Wang, R., Tian, Y., Wang, X., 2022. An attention based dual learning approach for video captioning. Applied Soft Computing 117, 108332.

Jishan, M.A., Mahmud, K.R., Al Azad, A.K., Rashid, M.R.A., Paul, B., Alam, M.S., 2021. Bangla language textual image description by hybrid neural network model. Indonesian J. Electr. Eng. Comput. Sci. 21 (2), 757–767.

Kamal, A.H., Jishan, M.A., Mansoor, N., 2020. Textmage: The automated bangla caption generator based on deep learning. pp. 822–826.

Khairullah, M., 2019. A novel steganography method using transliteration of bengali text. J. King Saud Univ.- Comput. Informat. Sci. 31 (3), 348–366.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization, CoRR abs/1412.6980.

Lin, C.-Y., 2004. Looking for a few good metrics: Rouge and its evaluation. In: Ntcir workshop.

Lin, K., Gan, Z., Wang, L., 2020. Multi-modal feature fusion with feature attention for vatex captioning challenge.

Lin, K., Li, L., Lin, C.-C., Ahmed, F., Gan, Z., Liu, Z., Lu, Y., Wang, L., 2022. Swinbert: End-to-end transformers with sparse attention for video captioning. pp. 17949–17958.

Luong, M.T., Pham, H., Manning, C.D., 2015. Effective approaches to attention-based neural machine translation. Association for Computational Linguistics, 1412–1421.

Mukta, M.S.H., Islam, M.A., Khan, F.A., Hossain, A., Razik, S., Hossain, S., Mahmud, J., 2021. A comprehensive guideline for bengali sentiment annotation, ACM Trans. Asian Low-Resour. Lang. Inf. Process. 21 (2), 1–19.

Palash, M.A.H., Nasim, M.A.A., Saha, S., Afrin, F., Mallik, R., Samiappan, S., 2002. Bangla image caption generation through cnn-transformer based encoder-decoder network. pp. 631–644.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. Bleu: a method for automatic evaluation of machine translation, in. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318.

Raj, A.H., Seum, A., Dash, A., Islam, S., Shah, F.M., 2021. Deep learning based video captioning in bengali. In: 2021 26th International Conference on Automation and Computing (ICAC), IEEE, pp. 1–6.

Shah, F.M., Humaira, M., Jim, M.A.R.K., Saha Ami, A., Paul, S., 2022. Bornon: Bengali image captioning with transformer-based deep learning approach. SN Comput. Sci. 3 (1), 1–16.

Siam, M., Valipour, S., Jagersand, M., Ray, N., 2017. Convolutional gated recurrent networks for video segmentation. In: 2017 IEEE International Conference on Image Processing (ICIP), IEEE, pp. 3090–3094.

Singh, A., Singh, T.D., Bandyopadhyay, S., 2021. Attention based video captioning framework for hindi. Multimedia Systems, 1–13.

Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.-F., Wang, W.Y., 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4581–4591.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning, PMLR, pp. 2048–2057.

Yang, Z., He, X., Gao, J., Deng, L., Smola, A., 2016. Stacked attention networks for image question answering, in. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 21–29.

Yang, Y., Zhou, J., Ai, J., Bin, Y., Hanjalic, A., Shen, H.T., Ji, Y., 2018. Video captioning by adversarial lstm. IEEE Trans. Image Process. 27 (11), 5600–5611.

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A., 2015. Describing videos by exploiting temporal structure, in. In: Proceedings of the IEEE international conference on computer vision, pp. 4507–4515.