Research Article

Amharic Language Image Captions Generation Using Hybridized Attention-Based Deep Neural Networks

Rodas Solomon and **Mesfin Abebe**

Department of Computer Science and Engineering, Adama Science and Technology University, Adama, Ethiopia

Correspondence should be addressed to Mesfin Abebe; mesfinabha@gmail.com

This study aims to develop a hybridized deep learning model for generating semantically meaningful image captions in Amharic Language. Image captioning is a task that combines both computer vision and natural language processing (NLP) domains. However, existing studies in the English language primarily focus on visual features to generate captions, resulting in a gap between visual and textual features and inadequate semantic representation. To address this challenge, this study proposes a hybridized attention-based deep neural network (DNN) model. The model consists of an Inception-v3 convolutional neural network (CNN) encoder to extract image features, a visual attention mechanism to capture significant features, and a bidirectional gated recurrent unit (Bi-GRU) with attention decoder to generate the image captions. The model was trained on the Flickr8k and BNATURE datasets with English captions, which were translated into Amharic Language with the help of Google Translator and Amharic Language experts. The evaluation of the model showed improvement in its performance, with a 1G-BLEU score of 60.6, a 2G-BLEU score of 50.1, a 3G-BLEU score of 43.7, and a 4G-BLEU score of 38.8. Generally, this study highlights the effectiveness of the hybrid approach in generating Amharic Language image captions with better semantic meaning.

1. Introduction

Ethiopia is a country that is located in the east part of Africa. It is a nation with over 84 nations and nationalities. The Amharic Language is the official working language of the Federal Democratic Republic of Ethiopia [1, 2]. It is spoken by over half of the population and is also spoken in nations such as Eritrea, Canada, the United States, and Sweden [3]. Amharic is written using the Fidel or Abugida script, derived from the ancient Ge'ez language of Ethiopia [1].

The use of artificial intelligence techniques to generate image captioning becomes widespread in the past few years [4]. This task involves the generation of textual descriptions for images using computer vision and natural language processing techniques. While the accuracy of English Language image captioning models has improved, they still have semantic incorrectness issues. This study aims to build a deep learning-based Amharic Language image caption model using attention mechanisms so that it can be applied in various Amharic Language software applications such as tools for the visually impaired, editing software, virtual assistants, and image searching [5].

The popularization of the deep learning approaches has enabled us to solve complex problems easily and successfully. Recent image captioning studies are using the encoder-decoder architecture which is originally used in machine translation [6–8]. The encoder-decoder approach submits images to an encoder that converts the visual elements into a fixed-length vector, which are then decoded into a textual description [6]. Pretrained CNN models are used for the encoder while long short-term memory (LSTM) or gated recurrent unit (GRU) neural networks are commonly used for the language generation. However, the encoder-decoder method is limited in its ability to preserve all source information in the fixed-length vector, and the unidirectional LSTM decoder only preserves past information which leads to poor outcomes for long sequential data [9].

In this study, we aim to address the limitations of previous models [10, 11] by incorporating an attention mechanism that focuses on both visual and linguistic features. The attention mechanism allows the model to extract only the relevant information, while also emphasizing high-level semantic features that better describe the image content [12]. Additionally, we employ a Bi-GRU architecture, which captures input information from both forward and backward directions. This enhances the capturing of both visual and textual features that can minimize the gap between them and lead to semantically richer image captions.

This study proposes a hybridized attention-based approach for Amharic Language captions to address the existing gaps in the area. The approach combines a Bi-GRU with an attention decoder to generate words by focusing on the required information that lead to semantically correct image captions. The Flickr8K and BNATURE datasets were used to evaluate the performance of the proposed model.

The study is organized in the following manner: Section 2 covers related works; Section 3 outlines the proposed approach; Section 4 explain into model building; Section 5 specifies the experiments and methodology; Section 6 presents results and discussions; and finally, the conclusion is presented in Section 7.

2. Related Works

This section describes the major works on image captions which can be divided into three sections. The sections are retrieval-based, template-based, and deep neural networks (DNNs) techniques.

2.1. Retrieval-Based Approaches.
The retrieval-based method employs a searching method to encounter proper image descriptions [13]. This method retrieves relevant nominee descriptive phrases from a database. Then generate an intelligible caption based on the text descriptions of the retrieved similar image sets [14]. In general, the retrieval-based approach required to have extensive data that covers each possible query picture.

2.2. Template-Based Approaches.
The template-based methods have specified templates with some empty slots for the caption generation [15, 16]. The approach proposes a triplet of scene components (object, action, and scene) that load the template space for the captions. The triplets provide a general idea of what the image is to generate a caption. Few authors [17] suggest a quadruplet template that includes (nouns-verbs-scenes-prepositions).

The template-based approaches can generate grammatically correct descriptions compared to retrieval-based methods. However, this approach is rigid and cannot generate variable-length image descriptions [15, 18]. Hence, the result image captions lack naturalness compared with manually generated sentences.

2.3. DNNs Approaches.
DNNs are a type of artificial neural networks (ANNs) that consist of multiple hidden layers, allowing for the creation of more complex models capable of higher levels of abstraction. The papers [19, 20] discuss the use of ANNs in computing for prediction purposes. The study focuses on training ANNs using meta-heuristic algorithms to improve precision and determining neural network (NN) input coefficients. An integrated algorithm was used and compared to other algorithms such as ant colony and invasive weed optimization. The results showed that the proposed algorithm had better convergence with NN coefficients and reduced prediction error in the NN. ANNs are a key component of the deep learning, forming the foundation for designing and deploying complex DNNs that learn and predict on intricate data. A deep learning based cyberattack detection and classification technique was introduced for intelligent systems (FDL-CADIS) [21]. The technique transforms malware binary files into 2-dimensional images and uses a MobileNetv2 model and an ensemble of voting-based classifiers for classification. The results of the experimental analysis showed promising performance in detecting and classifying malware cyberattacks. The image caption model leverages the power of DNNs, which relies on ANNs as a fundamental component to learn and make predictions on complex data and generate concise textual descriptions of the visual content.

DNNs dominate in the previous image caption techniques compared to that of template-based and retrieval-based methods. A recent study [19] proposes an innovative approach to image captioning using DNN architecture. The DNN employs a CNN as an encoder to extract image features which is then projected into a LSTM model. The authors also introduce a novel decoder neural network language model called structure-content neural language model (SC-NLM), which generates words using a combination of vector content and structure. The approach enhances the accuracy of the image captioning by leveraging the strengths of both the content and the structural information. Researchers have made several enhancements to the image caption model in order to produce semantically accurate and fluent captions. The problem of image caption generation using a custom ensemble model consisting of an Inception model and a 2-layer LSTM model was used in [22]. The results are evaluated using Bilingual evaluation understudy (BLEU) scores and it achieved a BLEU-4 score of 55.8%.

Reference [23] proposed utilizing semantic representation to improve image captioning by incorporating important image elements that are not captured by global feature representations. The authors employed region-based convolutional neural networks (R-CNN) to identify such elements, generating detailed captions that utilized semantic embedding. However, the static representation of semantic elements failed to consider the relevant image features. The paper [24] is used an area attention-based encoder-decoder model that associates parts of the image with the words of a description. In other study [25], the authors propose a novel method for generating image captions that combines spatial and channel-wise attention mechanisms over a 3D CNN features map. Unlike the previous approaches [24], which mainly used features from spatial locations, the authors

incorporate features extracted from different channels and multiple layers to focus on essential regions of the image. However, their approach overlooks the relevance of the sentence generation for image description.

The study [25] purposes a deep learning model that generates captions using a custom ensemble of LSTM and CNN algorithms. The author employs GRU and bi-directional LSTM for caption generation and uses Global Vectors (GloVe) embedding to generate the word vectors. This model does not include an attention mechanism. The study [26] is a variant model based on LSTM that was inspired by stimulus-driven and concept-driven attention mechanisms in psychology. The attention mechanism is used to detect image features and to obtain attention distribution in the images using a Gaussian filter applied to change region impact factors.

A question-answering model is proposed [27] using target relationship detection to answer questions about an image content. It incorporates a new attention mechanism and theories related to word vector space to improve image semantic tasks. The model uses question-based attention and converts target semantic information into word vector space to improve its generalization. The authors [28] proposed a semantic text summarization method of long videos. The researchers used unidirectional LSTM to generate the final caption. However, the unidirectional LSTM problem is restricted to have only past information which yielding poor outcomes for long sequences.

The aforementioned studies and most of the image captions studies are done on the English Language. There are a few studies, in other languages for example [29] which studied image caption in the Bangla Language. The authors used hybrid encoder-decoder architecture to generate the image caption. They, also create their own dataset for the study. This dataset is called Bangla natural language image to text (BNLIT) and contains 8700 images with single (one) annotation for each image.

Another study, [30] proposed an encoder-decoder model for image caption in the Bangla Language. The authors used Inception-v3 as a CNN encoder to extract the visual elements. They applied Bi-GRU to generate the textual description of the images. The authors argue that GRU has better results compared to LSTM. They also proposed a new dataset called BNATURE. This dataset is prepared based on Flickr8k dataset. The dataset contains of 8000 images with five Bengali Language captions for each image.

In general, the previous studies [29, 30] achieved good results in BLEU scores with Bilingual Language. However, their approach has limitation with the image feature extraction part. The approach directly passes the entire features into the language model without filtering the relevant features which comes from the image.

3. Proposed Approach

Figure 1 shows the proposed model architecture of the image caption generation for the Amharic Language. It comprises four main components that are discussed in detail in the next sections.

3.1. Word Embedding.
In NLP, word embeddings are used to convert words or document vocabulary into numerical form. The embeddings are applied to clean caption data obtained from image descriptions in training datasets. This involves transforming the sentences into word tokens. In this context, an input sentence containing T words is represented with $\{x_1, x_2, \ldots x_T\}$ tokens. Each word in the sentence, x_i, is transformed into a feature vector, e_i, through a matrix-vector product [31].

The weight matrix for word embedding denoted as W^{wrd}, belongs to the real number set $\mathbb{R}^{d^w}|V|$. Where d^w is the size of the word embedding (a hyperparameter) and $|V|$ is the size of the vocabulary (the number of distinct words in the corpus). Each word vector v^i is also given as $|V|$. The final output of the word embedding process is a real-valued feature vector, $\mathrm{Emb}_s = \{e_1, e_2, \ldots e_T\}$, which represents the input sentence $S = \{x_1, x_2, \ldots x_T\}$.

The formula for the conversion of a single word x_i into a feature vector e_i is

$$e_i = W^{\mathrm{wrd}} v^i, \tag{1}$$

where W^{wrd} is the weight matrix for word embedding and v^i is the one-hot encoded vector for the i-th word in the vocabulary.

3.2. CNN Image Encoder.
The image encoder uses the Inception-v3 model to extract local image features by removing the last two layers to obtain a global average pooling layer of $8 * 8 * 2048$. The pooling layer flattens the 8×8 feature map into a $64 * 2048$ vector. The final output is (L) feature vectors, each with a D-dimensional representation that corresponds to a portion of the image. (A) is an array of (L), where each vector (a_i) is a D-dimensional representation and belongs to $\mathbb{R}^{D\,(2048)}$.

$$A = [a_1, a_2, \ldots a_L], a_i \in \mathbb{R}^{D\,(2048)}. \tag{2}$$

Finally, the feature vectors are fed into a dense layer with ReLU activation which reduces the dimension to match the size of the word embedding Emb_s. The resulting feature vector is then used for visual attention to determine the relevant image section for generating captions.

3.3. Visual Attention Mechanism.
The visual attention function [30] is calculated in the proposed method by taking the hidden state ($h_{(t-1)}$) and the output of the encoder (a_i). The attention score is computed for each time step (t) and location (i) in the image by applying a nonlinear activation function (tanh) on ($h_{(t-1)}$) and (a_i), and then using the SoftMax activation function to get the attention distribution. Each part of the image is assigned a weight which represents its importance in combining feature vector (a_i). The visual feature vectors are 2048-dimensional vectors in the R space and part of the set $A = [a_1, a_2, \ldots a_L]$. Mathematically, the attention score is represented by α_{ti}

$$\alpha_{ti} = \mathrm{SoftMax}\left(W_a{}^T \tan h\left(W_a a_i + W_h h_{(t-1)}\right)\right), \tag{3}$$
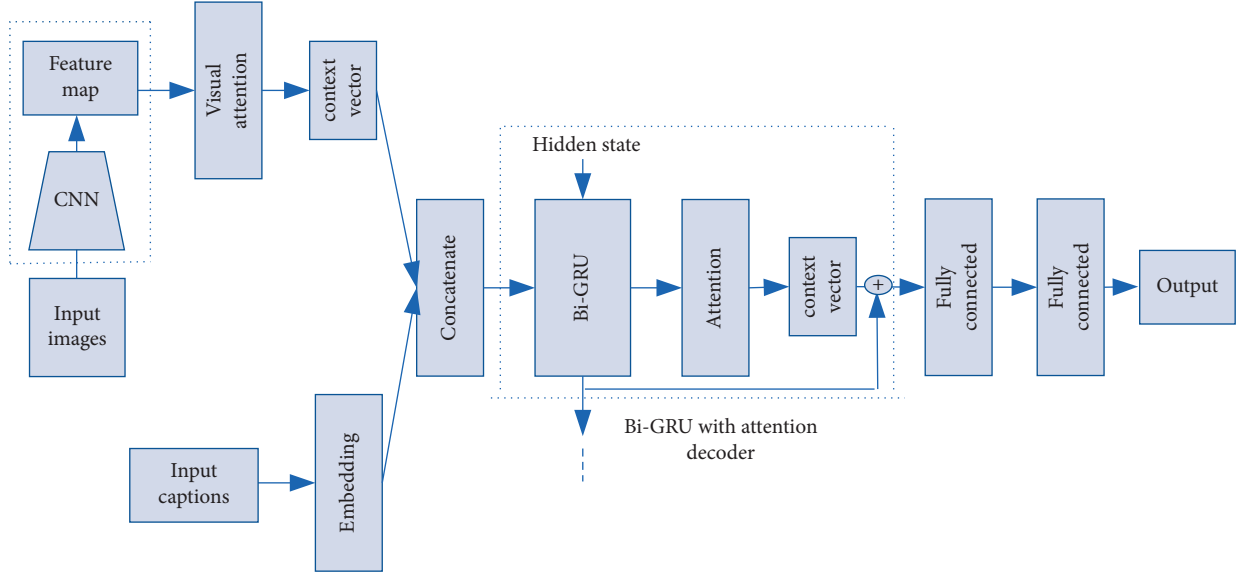
FIGURE 1: Architecture of the proposed model.

where W_a and W_h are weight matrices to learn the importance of i-th element in input sequence and the previous hidden state, respectively. The superscript T denotes the transpose of the weight matrix. The context vector is obtained by computing a weighted sum of the visual feature vectors using the attention weights (score) obtained from the SoftMax activation function. In other words, for each time step (t), the context vector represents the combined visual representation of each word in the input sequence, where the contribution of each word to the context vector is determined by its corresponding attention weight. Mathematically, the context vector (C_t) is expressed as

$$C_t = \sum_{i=0}^{L} \alpha_{ti} a_i. \tag{4}$$

This helps to reduce the gap between the image and the candidate captions for the image.

Finally, context vector output, and text feature representation outputs are combined to reduce the gap between the image and candidate caption. The result of this combination is represented by Con_x, which is calculated as the element-wise sum ($\oplus$) of the word embedding (Emb_s) and the context vector (C_t). This integration helps in combining the context vector and the word embedding information to produce the final output Con_x, which represents the combined representation of the image and text features. Con_x is represented as

$$Con_x = [Emb_s \oplus C_t]. \tag{5}$$

3.4. Bi-GRU with Attention Mechanism Language Decoder. The decoder in the Bi-GRU language decoder network takes the output from the concatenation layer as input and inputs it into a Bi-GRU network. The Bi-GRU computes both the forward and backward hidden sequences (represented as $\overrightarrow{h}$ and $\overleftarrow{h}$, respectively) is shown

in the Bi-GRU layer section in Figure 2. The computation is done as follows:

   (i) The update gate, z_t, is calculated using a sigmoid activation function (σ) on the weighted sum of the previous hidden state ($h_{(t-1)}$) and the concatenation of the visual attention context and text feature at time step t (Con_{xt}) and the corresponding bias term b_z.

   (ii) The reset gate, r_t, is calculated using a sigmoid activation function on the weighted sum of the previous hidden state ($h_{(t-1)}$) and the concatenation of the visual attention context and text feature at time step t (Con_{xt}) and the corresponding bias term b_r.

   (iii) The current memory content, $\widetilde{h}$, is calculated using a ReLU activation function (σ) on the weighted sum of the element-wise product of the reset gate output and the previous hidden state ($r_t * h_{(t-1)}$). The concatenation of the visual attention context and text feature at time step t (Con_{xt}) and the corresponding bias term b.

   (iv) The final memory at the current unit, h_t, is calculated as the element-wise sum of the previous hidden state with a weight of ($1 - z_t$) and the current memory content with a weight of z_t.

$$
\begin{aligned}
z_t &= \sigma\left(W_z \cdot \left[h_{(t-1)}, Con_{xt}\right] + b_z\right), \\
r_t &= \sigma\left(W_r \cdot \left[h_{(t-1)}, Con_{xt}\right] + b_r\right), \\
\widetilde{h} &= \sigma\left(W \cdot \left[r_t * h_{(t-1)}, Con_{xt}\right] + b\right), \\
h_t &= (1 - z_t) * h_{(t-1)} + z_t \widetilde{h}_t.
\end{aligned}
\tag{6}
$$

W, W_z, and W_r are the weight matrices for the current unit, update gate, and reset gate, respectively. b, b_z, and r_t are the corresponding bias terms for each gate. The model has
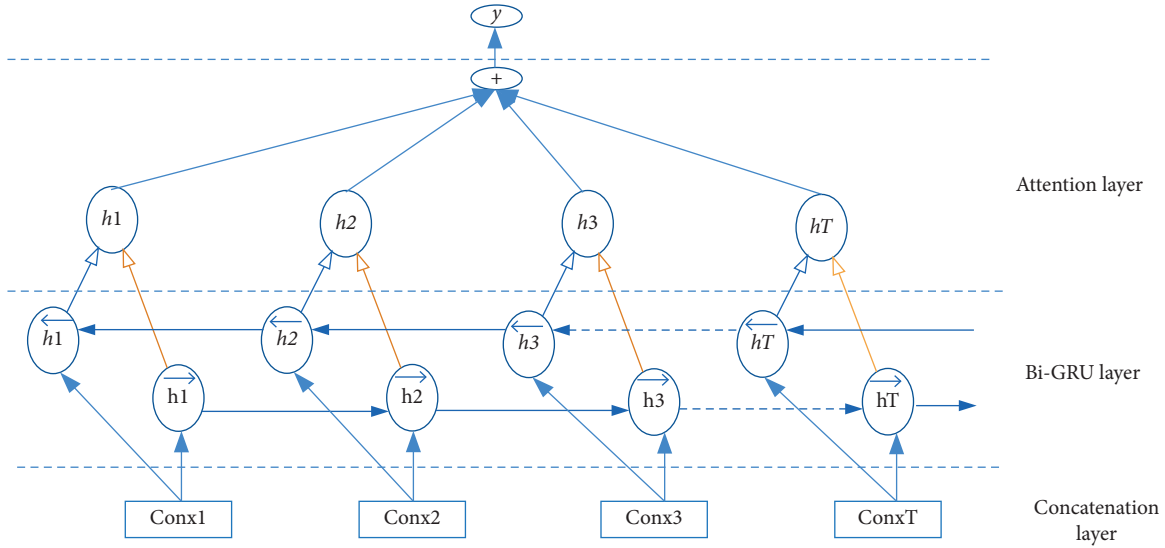
FIGURE 2: Bidirectional GRU with attention mechanism.

separate right and left layers that process the context in order. The following equation represents the final output of the jth word in the sequence. The forward and backward hidden sequence, $\overrightarrow{h}$ and $\overleftarrow{h}$, respectively, are concatenated (element-wise sum) to produce the final output h_j. The final output is the result of combining information from both the forward and backward passes of the Bi-GRU network.

$$h_j = \left[\overrightarrow{h_i} \oplus \overleftarrow{h_i} \right]. \tag{7}$$

3.5. Addictive Attention. The addictive attention describes a method for calculating attention context vectors in a language decoder. The method operates in two steps. The first step involves calculating a matching score (alignment score) e_{tj} between the current hidden state h_t and previous hidden state h_j using an additive projection. The second step involves calculating the attention context vector C_j for hidden state h_j based on the output of the alignment scores. The alignment scores are calculated using the following equation:

$$e_{tj} = V_a{}^T \tan h(W_a.h_t + U_a.h_j), \tag{8}$$

where V_a, W, and U_a are learned attention parameters and d is the dimensionality of the hidden state. The attention layer section in Figure 2 shows the calculation of the attention context vector, which is obtained by summing the product of alignment scores and hidden state h_j.

$$C_j = \sum_{j=1}^{T_x} e_{tj} h_j. \tag{9}$$

The final output of the addictive attention layer is transformed into a 1-dimensional representation by using a flattened layer. This representation is used as input for the final two dense layers. The first dense layer uses the

length parameter and the second dense layer outputs a probability distribution over the vocabulary to generate the final image caption. The use of a Bi-GRU network with an additive attention language decoder enhances the existing work to generate realistic captions by considering both the visual and textual relevant features, as shown in Figure 3.

The above figure displays the pipeline of both the baseline and proposed model. The solid black line represents the baseline model while the solid green line represents the proposed model. The red dotted line highlights the modifications and the gaps addressed by the proposed model in comparison to the baseline model.

The baseline model has two major limitations: (1) A direct connection between the encoder and the Bi-GRU decoder hinders the decoder from focusing on the crucial image information. (2) The baseline model lacks the ability to effectively select important visual-textual features during image caption generation. To address these limitations, the proposed model incorporates an additive mechanism into the language decoder that enables it to concentrate on the relevant visual-textual features.

4. Methodology

4.1. Dataset Preparation. In this study, two well-known datasets, Flickr8k (8,000 images), and BNATURE (8,000 images), were used for training and testing the proposed model. The datasets were divided into three sections with 75% of images for training, 12.5% for validation, and 12.5% for testing. Each image in the datasets was annotated with five separate captions in both English and Amharic Languages. The English captions were translated into Amharic using Google translate then reviewed by Amharic Language experts to correct grammar and semantic errors. A sample of the dataset is presented in Figure 4. Each image in the dataset has five English Language and Amharic Language captions.
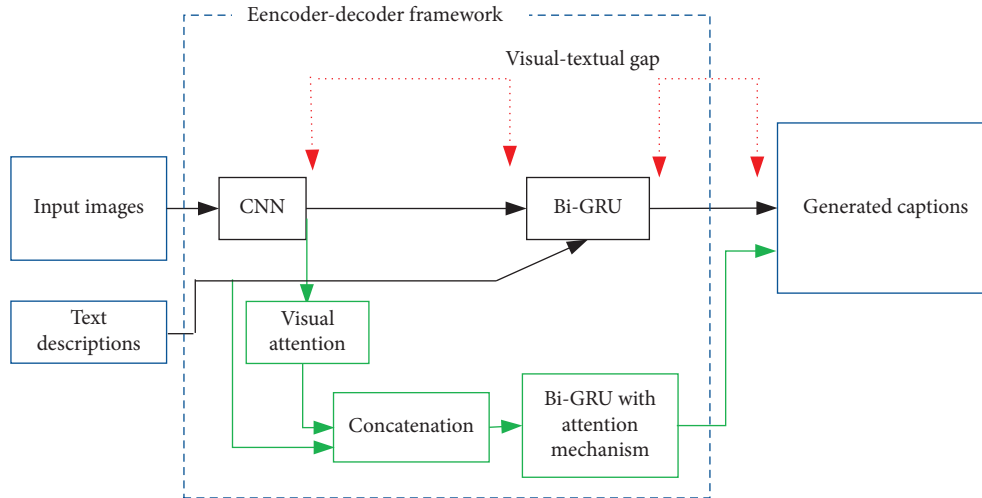
Figure 3: Modification of the existing approach.

Image	Image Captions
	(i) *A couple of several people sitting on a ledge overlooking the beach.* (ii) *A group of people sit on a wall at the beach.* (iii) *A group of teens sit on a wall by a beach.* (iv) *Crowd of people at the beach.* (v) *Several young people sitting on a rail above a crowded beach.*
	(i) *A black and white dog is running in a grassy garden surrounded by a white fence.* (ii) *A black and white dog is running through the grass.* (iii) *A Boston terrier is running in the grass.* (iv) *A Boston Terrier is running on lush green grass in front of a white fence.* (v) *A dog runs on the green grass near a wooden fence.*

Figure 4: Sample of the dataset with English Language captions.

The motivation of using the Flickr8k and BNATURE dataset has two reasons. First, their availability and suitable size for training with a low-power graphics processing unit (GPU), making them easier to analyse compared to other datasets such as Flickr30k and Microsoft common objects in context (MSCOCO). Second, these datasets captions were collected by human experts from Amazon Mechanical Turk (AMT) workers. The approach used precise instructions to verbalizing the main action depicted in the images.

4.2. Text Preprocessing Technique. Preprocessing is a crucial step in the development of machine learning and deep learning models. It aims to convert the raw data into a format that is suitable for processing and analysis. This step is especially critical when working with Amharic Language, as they require specific language-based rules and the removal of irrelevant words and characters.

The data cleaning process is applied to ensure that the captions are in a standardized format to make it easier for the algorithms. This practice includes the lower-casing of words to avoid duplications and the removal of special characters, such as "+," "%," "$," "#," "@," etc. Additionally, any words that contain numbers, like "Hello123," are also removed. To further enhance the accuracy of the captions, any inappropriate words and symbols are manually adjusted to ensure that the translated captions maintain their original meaning.

In this study, data cleaning was performed on a dataset of 40,000 (8000 ∗ 5) sentences. Once the data cleaning was complete, each caption sentence was marked to indicate the beginning and the end using the tags "<start>" and "<end>." This was done to indicate the start and end of the image captions for the algorithms.

After the data cleaning process, the next steps involved in the preprocessing stage are tokenization and vectorization. Tokenization is the process of breaking down the caption sentences into smaller units, such as words, characters, or sentences. In this study, the image captions were broken down into word-level tokens. The vectorization step involves converting the tokenized captions into numerical representations, making it easier for the algorithms to process the information.

4.3. Image Preprocessing. The image dataset used in this study consists of 8,000 unique images that have been preprocessed to ensure uniformity in images size. The images were resized to a resolution of 299 width, 299 height, and 3 colour channels. In addition to scaling, grayscale histogram and data augmentation techniques were applied to further preprocess the image datasets.

To extract the image features, a pretrained CNN model was used instead of training the model from scratch. Training a convolutional network from scratch requires a large dataset and high computing power; therefore the Inception-v3 CNN model is used for this purpose. Inception-v3 is a pretrained model that was originally trained for image classification tasks and has a lower number of parameters compared to other pretrained CNN models such as VGG-16 and ResNet-50.

The Inception-v3 model is used to extract the most important features from the images as a feature extraction technique. The feature extraction was implemented by removing the last SoftMax layer of the Inception-v3 model and focusing on the 2048 features of each image. This approach ensures the capturing and utilization of the most significant information in the deep learning models building process. As a result, the model can give more accurate and reliable image captions.

4.4. Evaluation Metric. BLEU is a Precision-Based Metric for Machine-Generated Text. Evaluating the quality of a machine-generated text is crucial in the development of language models. The BLEU metric is a widely used method for measuring the accuracy of machine-generated text. It is a precision-based metric that ranges from 0 to 100, where 100 indicating a perfect match and 0 indicating a perfect mismatch.

This approach evaluates the quality of machine-generated text by comparing the *n*-grams of the predicted outcome of a model to the n-grams of the actual data [32]. A high BLEU score, close to 100, indicates a better model, while a score close to zero is considered as a poor model. BLEU has been widely used in the field of natural language processing and has been found to be a reliable and effective method for evaluating the quality of machine-generated text.

5. Experiments and Model Building

In this study, two experiments are conducted to evaluate the performance of the proposed model in comparison to the base model. The first experiment was performed on the CNN-Bi-GRU encoder-decoder model, while the second experiment is performed on the proposed hybridized attention-based CNN-Bi-GRU model. Experiment on the base model includes:

(1) Image features were extracted using the pretrained Inception-v3 model

(2) The extracted features are fed into the CNN encoder

(3) The output of the CNN encoder is combined with the word embedding layer and sent to the Bi-GRU layer

(4) The final output of the model is the predicted word probability based on the Bi-GRU unit.

The proposed hybridized attention-based CNN-Bi-GRU model experiment covers:

(1) Attention mechanisms are implemented on both the CNN encoder and the Bi-GRU decoder

(2) The first visual attention was placed between the CNN and the Bi-GRU to allow the Bi-GRU to focus only on the relevant image features

(3) Extending the attention mechanism to the existing Bi-GRU decoder aimed to reduce the gap between the vision and the textual component and generate semantically correct image captions.

The experiments were performed using the Flick8k dataset in both English and Amharic languages caption. The result of the experiments is evaluated using the BLEU scores. The model is training and optimized with the hyperparameters shown in Table 1. Generally, the study is conducted based on the following assumptions:

(i) The image features that are extracted using the pretrained Inception-v3 CNN model will provide enough information for the encoder-decoder models to generate image captions

(ii) The tokenization and vectorization techniques applied to the image captions can enable the models to effectively comprehend language patterns

(iii) The hybridized attention mechanism, applied to both the CNN encoder and the Bi-GRU decoder, can enhance the performance of the model and generate semantically correct image captions.

On the other hand, it is important to note that the result of the model is influenced by the limitations in the data and algorithms used.

6. Results and Discussion

The experimental results show that the proposed hybridized model outperforms the basic model in terms of image captioning accuracy for the Amharic Language. As indicated in Table 2, the hybridized model achieves a 15%, 12%, 10%, and 10% improvement in 1G-BLEU, 2G-BLEU, 3G-BLEU, and 4G-BLEU scores, respectively. Gram (G) refers to the number of words in the caption. This significant increase in performance can be attributed to the integration of visual attention and Bi-GRU with an attention mechanism.

The visual attention allows the model to focus on the most important parts of the image while the Bi-GRU attention mechanism selects the most relevant words to describe the content of the image. The results of the captions are more descriptive and accurately reflect the context of the image. Furthermore, the use of Bi-GRU with an attention decoder during the caption generation process ensures that the generated words are highly relevant and appropriate for the image context.

The results of the two models are compared on the BNATURE and Flickr8k datasets, as show in Table 3. The purpose of these experiments is to evaluate the robustness and generalizability of the models using different datasets. The results with the BNATURE dataset provide more evidence on the performance improvement of the proposed hybridized model compared to the base model. The implication of these results lies in their demonstration of the ability of the hybridized model to adapt to different data sources without losing its sematic accuracy. This indicates that the proposed model is not only effective on the Flickr8k dataset but also on other similar datasets, making it a more flexible and practical solution for image captioning. The proposed model generated image captions with better meaning compared to the basic model.

TABLE 1: Hyperparameters setup.

Hyperparameter	Values
Batch size	32
Embedding dimension	256
Dense layer	256
Hidden layer (unit)	512
Dropout for encoder and decoder	0.2 (for dense layer) and 0.2 (for GRU cell)
Recurrent dropout	0.2
Recurrent activation	ReLU
Kernel regularize and recurrent regularize	$L2$ (0.001) and $L2$ (0.1)
Optimizer	Adam
Epoch	35

TABLE 2: BLEU score result of the models using the Amharic captions Flickr8k dataset.

Model	1G-BLEU	2G-BLEU	3G-BLEU	4G-BLEU
Base model	46.4	38.2	33.0	28.5
Hybridized model	**61.6**	**50.1**	**43.7**	**38.8**

The results of the proposed model demonstrate its dominance over other models, as it achieved better results on all BLEU scores in both datasets. The model's training accuracy on the BNATURE dataset was 0.925, with a testing accuracy of 0.928. The loss during the training phase was approximately 0.186 and during the testing phase, it was 0.181, as shown in Figure 5. Similarly, on the Flickr8k dataset, the training accuracy was 0.882 and the testing accuracy was 0.885. The overall training loss was 0.303 and the test loss was 0.295, as demonstrated in Figure 6.

One of the advantages of using the Bi-GRU technique in this study is its simple configuration compared to Bi-LSTM. It only uses two gates, the update and reset gates, which results in faster speed calculation. Moreover, incorporating the DNNs techniques such as Inception-v3 and Bi-GRU improves the quality of the image captions. The Inception-v3 was used as the encoder to obtain visual features, while the decoder Bi-GRU was used to predict the words that make up the image caption. The results of the study show that the use of attention-based Bi-GRU or the hybridized model results in a better image captions.

The proposed model has demonstrated its advantages over other models in terms of image captioning accuracy and speed of computation. The results of the study show that the use of Bi-GRU and attention-based techniques significantly improves the semantic quality of the image captions.

However, there are also some shortcomings of the proposed model that need to be considered. One of the main drawbacks is the increased complexity of the model. The use of multiple components such as Inception-v3 and Bi-GRU with attention mechanisms results in a more complex model that requires more computational resources and longer training times. Additionally, the model may not perform well on images with complex or unusual content, as it relies on the accuracy of the visual features extracted by the Inception-v3 encoder. In such cases, alternative encoders or

TABLE 3: The models performance on Flickr8k and BNATURE datasets.

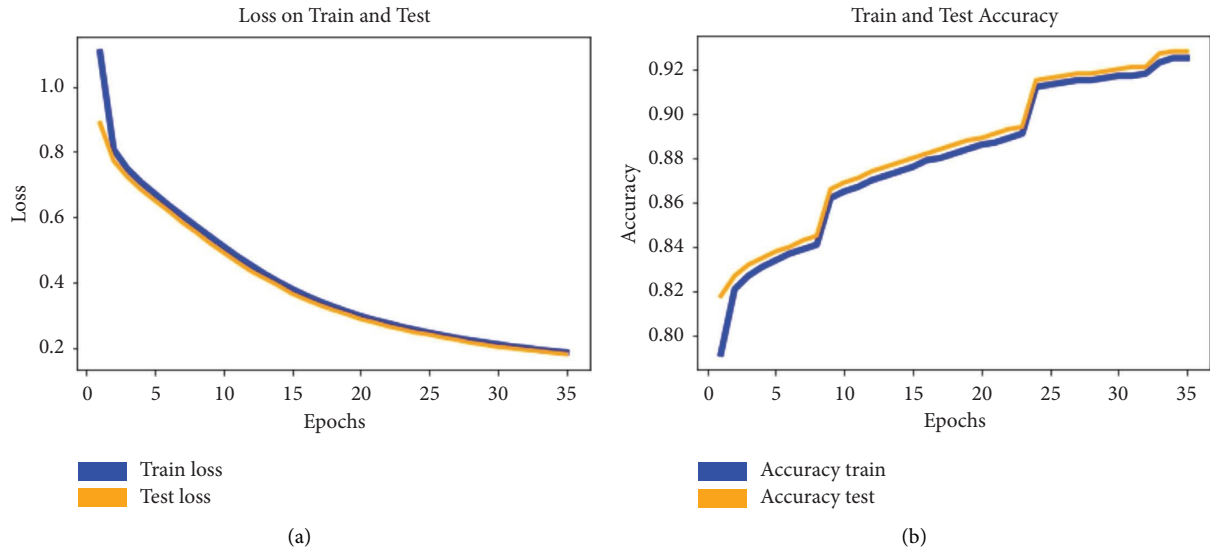| Dataset	Model	1G-BLEU	2G-BLEU	3G-BLEU	4G-BLEU
BNATURE	CNN-Bi-GRU	42.6	27.9	23.6	16.4
	Hybridized model	**63.1**	**53.3**	**47.4**	**42.7**
Flickr8k English	Bag-LSTM	59.7	41.4	28.1	18.2
	Hybridized model	**65.6**	**53.2**	**45.3**	**39.4**

FIGURE 5: Performance of the model on the BNATURE dataset. (a) Loss on train and test. (b) Train and test accuracy.
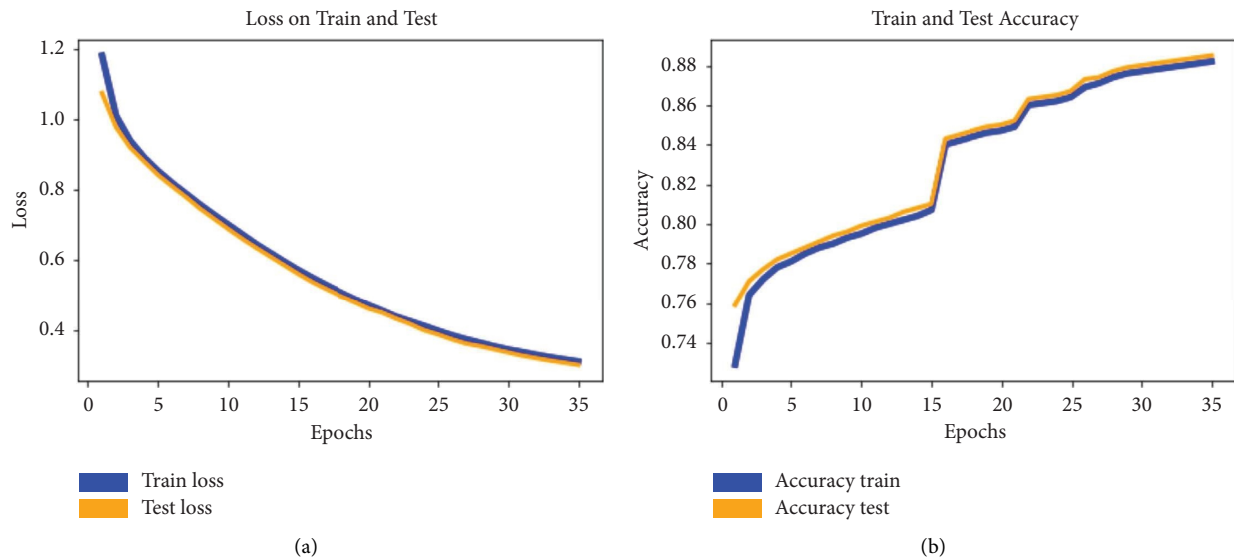

FIGURE 6: Model performance on the Flickr8K dataset. (a) Loss on train and test. (b) Train and test accuracy.

more advanced techniques may need to be employed to improve performance.

Another shortcoming is that the model may not generalize well to different languages or cultures, as it is trained on a specific dataset that represents a particular language and cultural context. To address this issue, future research could explore the use of multilingual or cross-cultural datasets to train the model and increase its generalizability.

In summary, while the proposed model has demonstrated its effectiveness in improving the quality of image captions, it also has its limitations that need to be considered in future research. Despite these limitations, the results of
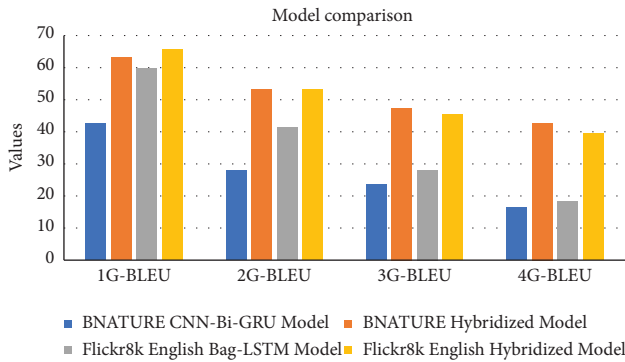
Figure 7: Model comparison.

the study provide strong evidence of the potential of using Bi-GRU with attention mechanisms for image captioning.

The proposed model improves image captioning by focusing on essential image features and corresponding text features. Results show a 10% increase in performance, as measured by the 4G-BLEU score, compared to the basic Bi-GRU decoder.

The experimental results in Table 3 reveal the performance of the proposed hybridized model using two different datasets. To assess the effectiveness of the proposed approach, the results were compared with two base models, CNN-Bi-GRU [30] and Bag-LSTM [7].

The results showed that the proposed model outperforms both baseline models on all four BLEU scores for both datasets. On the BNATURE dataset, the proposed model achieved a 20%, 25%, 24%, and 21% improvement in 1G-BLEU, 2G-BLEU, 3G-BLEU, and 4G-BLEU scores, respectively, compared to the CNN-Bi-GRU model. On the Flickr8k dataset, the improvement was 6%, 12%, 17%, and 21% higher than the Bag-LSTM model on 1G-BLEU, 2G-BLEU, 3G-BLEU, and 4G-BLEU scores.

The integration of visual attention and Bi-GRU with an attention mechanism decoder has proven to be a more effective approach for generating image captions. The hybridized model has demonstrated a significant improvement in performance compared to the existing models, with a 21% increase in 4G-BLEU score compared to both CNN-Bi-GRU and Bag-LSTM. The results in Figure 7 further confirm the superiority of the proposed hybridized model in generating high-quality image captions.

7. Conclusion

The significance of Amharic image captioning for a range of Amharic language-based applications has been highlighted in this study. By combining the image processing and text processing domains, the challenge of generating grammatically and semantically correct captions has been addressed. A hybridized attention-based CNN-Bi-GRU model has been proposed to overcome these challenges and enhance the quality of Amharic image captions.

The proposed model comprises of four main components, including the word embedding, image encoder, the visual attention mechanism, and the language decoder.

Word embedding is a technique in NLP that maps each word in a vocabulary to a high-dimensional vector of real numbers which can be used as a representation of the word's meaning. The image encoder extracts image feature. The visual attention mechanism focuses on the critical areas of the image and the language decoder learns a two-way (bidirectional) long-term dependency between sequential information to produce an image description. Experiments on the translated Flickr8k and BNATURE datasets have shown that the proposed model outperforms the baseline CNN-Bi-GRU and Bag-LSTM models.

The results indicate that integrating the visual attention mechanism and the Bi-GRU language decoder into the image captioning process improves the semantics of the generated descriptions. The study concludes that this approach to Amharic image captioning is an effective means of improving the quality of image descriptions and is a valuable contribution to the field. In the future, we aim to further develop our dataset to be more similar to the Flickr8k English dataset by collecting images that reflect the diverse cultures of our country. This will involve collecting images from various sources and annotating each image with five Amharic sentences. Additionally, we plan to utilize the successful proposed model in a new task such as recognizing activities in Amharic videos.

Data Availability

The data used to support this study are available in "https://drive.google.com/file/d/1geLObzMiTaJLpOwE1MMzZ4BXc64Go3g2/view?usp=sharing."

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

References

[1] A. A. Bekele, "Automatic generation of Amharic math word problem and equation," *Journal of Computer and Communications*, vol. 8, no. 8, pp. 59–77, 2020.

[2] T. Fikre and A. Ababa, *Effect of Preprocessing on Long Short Term Memory Based Sentiment Analysis for Amharic Language*, Addis Adaba University, Addis Ababa, Ethiopia, 2020.

[3] A. A. Kesito, "Character recognition of bilingual amharic-latin printed documents," 2018, http://etd.aau.edu.et/handle/123456789/18541.

[4] C. Masotti, D. Croce, and R. Basili, "Deep learning for automatic image captioning in poor training conditions," *Italian Journal of Computational Linguistics*, vol. 4, no. 1, pp. 43–55, 2018.

[5] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5561–5570, Salt Lake City, UT, USA, June 2018.

[6] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," pp. 1–11, 2019, https://arxiv.org/abs/1906.05963.

[7] P. Cao, Z. Yang, L. Sun, Y. Liang, M. Q. Yang, and R. Guan, "Image captioning with bidirectional semantic attention-based guiding of long short-term memory," *Neural Processing Letters*, vol. 50, no. 1, pp. 103–119, 2019.

[8] S. Bai and S. An, "A survey on automatic image caption generation," *Neurocomputing*, vol. 311, pp. 291–304, 2018.

[9] D. M. Agughalam, "Bidirectional lstm approach to image captioning with scene features davis munachimso agughalam supervisors," National College of Ireland, Dublin, Ireland, MSc. Data Analytics, 2020.

[10] C. Wang, H. Yang, and C. Meinel, "Image captioning with deep bidirectional LSTMs and multitask learning," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 2, pp. 1–20, 2018.

[11] M. Humaira, S. Paul, M. Jim, A. S. Ami, and F. M. Shah, "A hybridized deep learning method for Bengali image captioning," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, pp. 698–707, 2021.

[12] X. Liu and Q. Xu, "Adaptive attention-based high-level semantic introduction for image caption," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 16, no. 4, pp. 1–22, 2020.

[13] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi, "Collective generation of natural image descriptions," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 359–368, Jeju Island, Korea, July 2012.

[14] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2Text: describing images using 1 million captioned photographs," in *Proceedings of the 25th Advances in Neural Information Processing Systems*, pp. 1–9, Greece, Balkans, December 2011.

[15] M. D. Zakir Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," 2018, https://arxiv.org/abs/1810.04020.

[16] A. Farhadi, M. Hejrati, M. Amin Sadeghi et al., "Every picture tells a story: Generating sentences from images," in *Proceedings of the 2010 European Conference on Computer Vision*, vol. 6314, pp. 15–29, Heraklion, Crete, Greece, September 2010.

[17] Y. Yang, C. L. Teo, H. Daumé, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 444–454, Doha, Qatar, July 2011.

[18] M. G. Martínez, A. B. Rué, and C. R. Jordi, "Deep learning for image captioning an encoder-decoder architecture with soft attention," 2019, https://openaccess.uoc.edu/bitstream/10609/100446/6/magomarTFM0619memory.pdf.

[19] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal Neural Language models," pp. 1–13, 2014, http://arxiv.org/abs/1411.2539.

[20] A. A. Movassagh, J. A. Alzubi, M. Gheisari et al., "Artificial neural networks training algorithm integrating invasive weed optimization with differential evolutionary model," *Journal of Ambient Intelligence and Humanized Computing*, vol. 6, 2021.

[21] J. A. Omar and A. Alzubi, "Fusion of deep learning based cyberattack detection and classification model for intelligent systems," *Cluster Computing*, vol. 26, 2022.

[22] J. A. Alzubi, R. Jain, P. Nagrath, S. Satapathy, S. Taneja, and P. Gupta, "Deep image captioning using an ensemble of cnn and lstm based deep neural networks," *IOS Press*, vol. 40, no. 4, 2021.

[23] X. Zhang, S. He, X. Song, R. W. Lau, J. Jiao, and Q. Ye, "Image captioning via semantic element embedding," *Neurocomputing*, vol. 395, pp. 212–221, 2020.

[24] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, "Areas of attention for image captioning," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, vol. 2, pp. 1251–1259, Venice, Italy, October 2017.

[25] L. Chen, H. Zhang, J. Xiao et al., "SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6298–6306, Honolulu, HI, USA, January 2017.

[26] S. Ding, S. Qu, Y. Xi, and S. Wan, "Stimulus-driven and concept-driven analysis for image caption generation," *Neurocomputing*, vol. 398, pp. 520–530, 2020.

[27] Y. Xi, Y. Zhang, S. Ding, and S. Wan, "Visual question answering model based on visual relationship detection," *Signal Processing: Image Communication*, vol. 80, Article ID 115648, 2019.

[28] S. Ding, S. Qu, Y. Xi, and S. Wan, "A long video caption generation algorithm for big video data retrieval," *Future Generation Computer Systems*, vol. 93, pp. 583–595, 2019.

[29] A. Jishan, "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.

[30] A. M. Faruk, H. A. Faraby, M. M. Azad, M. R. Fedous, and M. K. Morol, "Image to Bengali caption generation using deep CNN and bidirectional gated recurrent unit," in *Proceedings of the 2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pp. 19–21, Dhaka, Bangladesh, December 2020.

[31] P. Zhou, W. Shi, J. Tian et al., "Attention-based bidirectional long short-term memory networks for relation classification," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 207–212, Berlin, Germany, August 2016.

[32] B. B. Phukan and A. R. Panda, "An efficient technique for image captioning using deep neural network," 2020, https://arxiv.org/abs/2009.02565.