Image Caption Synthesis for Low Resource Assamese Language using Bi-LSTM with Bilinear Attention

Pankaj Choudhury[1], **Prithwijit Guha**[2], **Sukumar Nandi**[3]
[1]Centre for Linguistic Science and Technology
[2]Department of Electronics and Electrical Engineering
[3]Department of Computer Science and Engineering
Indian Institute of Technology Guwahati, Assam, India
{pankajchoudhury[1], pguha[2], sukumar[3]}@iitg.ac.in

Abstract

The task of Automatic Image Captioning (AIC) involves the synthesis of semantically and syntactically correct natural language descriptions for an input image. Most existing works on AIC have focused on caption generation in the English language. In contrast, very few efforts have been made to synthesize captions in Indian languages, especially the low-resource ones. The first contribution of this paper is the creation of two image caption datasets for Assamese language. This includes Flickr30K-Assamese Caption (Flickr30K-AC) and COCO-Assamese Caption (COCO-AC) dataset. These datasets are created by translating Flickr30K and MSCOCO English captions to Assamese. The semantic and syntactic errors in translated captions are further corrected manually. The manual correction was performed to preserve the linguistic characteristics of Assamese language for training caption generators. Second, a Bi-directional LSTM (Bi-LSTM) based model with bilinear attention is proposed for generating the Assamese captions. The model performance is evaluated through qualitative and quantitative measures (BLEU-n and CIDEr scores) and benchmarked against three baseline models.

1 Introduction

Automatic Image Captioning (AIC) refers to the task of composition of natural language description of an input image. It is an AI complete task involving complex interactions between image understanding and language modeling. Recent research in image captioning has witnessed significant performance gain owing to the advancements in vision and language problems formulated in deep learning framework. This enables real-life applications of image captioning, like assisting visually impaired persons and generating reports for general (e.g. indoor or outdoor scenes) or specific (e.g. medical or remote sensing) images (Stefanini et al., 2022).

The Assamese language is the official language of North-East Indian state of Assam with more than 15 million native speakers (Chandramouli and General, 2011). The Assamese language is a member of the Indo-Aryan language family and has similarities with Bengali, Odia, and Hindi. The present-day Assamese script evolved from the Brahmi script (Pathak et al., 2022) and is written from left to right (see Figure 1). Additionally, Assamese language adheres subject-verb-object (SVO) word order. Furthermore, Assamese grammar uses different type of identifiers to express the meaning for gender (e.g. এজন ল'ৰা *"ejon lora"* "A boy", এজনী ছোৱালী *"ejoni chowali"* "A girl "), shape of object (e.g. এখন টেবুল *"ekhon tebul"* "A table", এটা আপেল *"eta aapel"* "An apple"). Thus, the use of classifier word is dependent on the next word. The highly inflected form to convey person (e.g. মই ভাত খাওঁ *"moi vaat khao"* "I eat rice", আপুনি ভাত খায় *"aapuni vaat khai"* "You eat rice"), aspect, tense is a unique feature of Assamese language. Furthermore, Assamese grammar makes extensive use of honorifics and gender distinctions, highlighting social standing and respect. Despite its intricate grammar and distinctive linguistic features, the research on NLP remains unexplored for Assamese language due to limited resources.

Classical approaches to AIC used template and retrieval based techniques (Farhadi et al., 2010; Kulkarni et al., 2013; Li et al., 2011; Mitchell et al., 2012). The Template-based methods generates captions by filling blank spaces in a fixed template sentence. In con-

Vowels and vowel symbols

অ	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও	ঔ
	◌া	ি◌	◌ী	◌ু	◌ূ	◌ৃ	ে◌	ৈ◌	ে◌া	ে◌ৗ
[ɔ]	[aː]	[i]	[iː]	[u]	[uː]	[ri]	[e]	[oj]	[o]	[oʊ]

Consonants

ক	খ	গ	ঘ	ঙ	চ	ছ	জ	ঝ	ঞ
[k]	[k^h]	[g]	[g^h]	[ŋ]	[tʃ]	[tʃ^h]	[dʒ]	[dʒ^h]	[ɲ]
ট	ঠ	ড	ঢ	ণ	ত	থ	দ	ধ	ন
[ʈ]	[ʈ^h]	[ɖ]	[ɖ^h]	[ɳ]	[t̪]	[t̪^{h}]	[d̪]	[d̪^{h}]	[n]
প	ফ	ব	ভ	ম	য	ৰ	ল	ৱ	শ
[p]	[p^h]	[b]	[b^h]	[m]	[dʒ]	[r]	[l]	[ʋ]	[x]
ষ	স	হ	ক্ষ	ড়	ঢ়	য়			
[x]	[x]	[ɦ]	[k^{h}j]	[r]	[ɽ]	[j]			

Digits

০	১	২	৩	৪	৫	৬	৭	৮	৯
xuinno	ek	dui	tini	sari	pas	soy	xat	ath	no
Zero	One	Two	Three	Four	Five	six	seven	Eight	Nine

Figure 1: Assamese script along with phonetic notation (Samudravijaya, 2021).

trast, retrieval-based methods retrieved captions from a set of existing sentences according to visually similar images. The classical approaches could generate grammatically and syntactically correct captions. However, predefined templates were unable to create variable-sized sentences. Also, generated sentences often described irrelevant image contents. Recent approches to AIC are formulated in the deep learning framework (Bai and An, 2018). Recent works on incorporation of attention mechanism in these approaches have demonstrated improved performance (Xu et al., 2015; Anderson et al., 2018). However, AIC continues to be a challenging problem on account of the complexities of image content understanding, language generation and vision-language interactions (Holzinger et al., 2021).

The research on image captioning is mainly focused on English language due to the availability of large scale image-caption datasets. In contrast, very few works are dedicated to image caption generation for Indian languages, especially the low-resource Assamese language. This work is primarily motivated by the absence of a standard image caption dataset in Assamese language. The major contributions of this work are as follows

- Creation of an Assamese image caption dataset by translating the captions of Flickr30K and MSCOCO from English to Assamese using Microsoft Translator[1]. All semantic and syntactic translation errors are further corrected manually. This manual curation is performed to preserve the linguistic characteristics of the Assamese language.
- An Assamese image captioning model using Bi-LSTM and bilinear attention is developed to evaluate the Flickr30K-AC and the COCO-AC datasets. Observations from Assamese grammar indicate that the prediction of a certain word in image caption will often depend on the next word. This motivated the use of Bi-LSTM decoder which enables the captioning model to use forward and backward context. This in turn helps the model to understand linguistic properties like, use of classifiers, which is dependent on the next word. This is crucial to generate syntactically and semantically correct Assamese captions. Further, a bilinear attention module is employed to boost the performance of the proposed model.

2 Related Work

2.1 Image Captioning

Motivated by neural machine translation, the recent methods in AIC are formulated in the deep encoder-decoder framework (Sutskever et al., 2014). Most deep encoder-decoder

[1] https://azure.microsoft.com/

frameworks methods have used the CNN as visual feature encoder and the LSTM (or GRU) as word sequence (caption) decoder. Vinyals *et al.* (Vinyals et al., 2015) proposed the VGG16 for visual feature encoding and an LSTM as decoder. The visual features were fed to the LSTM as the initial hidden state. The most probable caption words were predicted based on current input and the previous hidden state. However, the visual information became weaker on account of the vanishing gradient problem while generating long captions. Accordingly, Xu *et al.* (Xu et al., 2015) integrated an *attention mechanism* into the encoder-decoder framework. Here, the attention weights were computed for the patches of the input image in each decoding stage. Anderson *et al.* (Anderson et al., 2018) proposed the use of *bottom-up features* derived from salient scene objects. As captions describe mostly salient objects and their inter-relation(s), bottom-up features minimize noise and computational time while enhancing model performance. Huang *et al.* (Huang et al., 2019) applied multi-head self attention to enhance visual features. Most attention mechanisms in image captioning exploit first order interaction between visual features and captions. Second order interaction was introduced by Pan *et al.* (Pan et al., 2020) through bilinear attention. This was performed through outer product between visual features and caption and provided comparatively richer representations. However, all these models employed unidirectional LSTM as decoder, which is limited by only the past context for current word generation. Wang *et al.* (Wang et al., 2016) proposed a Bi-LSTM that utilized both past and future context for caption generation. This method used two end-to-end trainable LSTMs that generated captions in both forward and backward directions with the help of global image features. Finally, the word sequence with the maximum probability was chosen as the final caption. Later, Cao *et al.* (Cao et al., 2019) proposed the bag-LSTM, which used Bi-LSTM with a semantic attention mechanism.

2.2 Image Captioning in Indian Languages

Most works in the literature have focused on caption generation in English language. In contrast, very few works exist on Indian languages like Hindi, Bengali, and Assamese. Rahman *et al.* (Rahman et al., 2019) created the BanglaLekha-ImageCaption dataset with 16K annotated images for Bengali language. Shah *et al.* (Shah et al., 2021) proposed a transformer based model for caption generation in Bengali language. Mishra *et al.* (Mishra et al., 2021a) translated MSCOCO English captions to Hindi and used an attention-based encoder-decoder model to generate captions in Hindi. Furthermore, Mishra *et al.* (Mishra et al., 2021b) proposed a transformer model to improve caption generation in Hindi. Mishra *et al.* (Mishra et al., 2022) employs a hierarchical RNN model for image paragraph generation in Hindi language. For Assamese language Das *et al.* (Das and Singh, 2022) proposed a new image captioning system using attention. Nath *et al.* (Nath et al., 2022) proposed an Assamese caption dataset by combining translated captions of MSCOCO and Flickr30K English captions to Assamese. However, the method proposed by Nath et al. *et al.* has several limitations. First, there is no indication of any curation process for the semantic and syntactic errors presents in the translated captions of the dataset utilized in their study. The dataset has been translated from English to Assamese and used for training. Further analyses of the translated dataset reveal that Assamese captions contain semantic and syntactic errors that must be handled. Second, the model employed in their study is based on the “Show, Attend, and Tell” (Xu et al., 2015) paper, the decoder is replaced from LSTM with GRU and uses VGG16 and EfficientNetB3 as the image encoders. However, the paper does not give any insights into how the results would change if LSTM is used as the decoder. Additionally, it is not explicitly stated which BLEU (BLEU-1, BLEU-2, BLEU-3, or BLEU-4) score was used to measure performance, which could have an impact on how the results are interpreted. Furthermore, the higher BLEU score for Flickr30K compared to COCO is unclear, given that the COCO dataset has significantly more images available. The methods used to determine the maximum permissible caption length and the minimum frequency for choosing vocabulary words are also not well explained in the study. Last, considering the popularity of beam search for improving caption quality, there is no mention of its usage. It would be beneficial to employ beam search for better caption generation in the Assamese language.

3 Assamese Image-Caption Dataset

The Flickr30K and MSCOCO image caption datasets are the most popular datasets in the field of AIC. The images of Flickr30K and MSCOCO dataset are collected from the website "Flickr.com". The dataset covers a wide variety of concepts such as people, animals, nature, buildings and more. The Flickr-AC and COCO-AC dataset is created by translating English captions from Flickr30K and MSCOCO-2017 dataset to Assamese using the Microsoft translator. Later, semantic and syntactic errors are manually corrected for the translated Assamese corpus. A few examples of the Flickr30K-AC and COCO-AC dataset is shown in Table 2 along with original English caption (C1), translated Assamese caption (C2) and Corrected Assamese caption (C3). Example in Table 2(a) shows the classifier এখন "*ekhon*" (A) is used with ফুটবল "*futbol*" (Football). However, the word ফুটবল is referring to a round object and use of এখন is syntactically incorrect. During the curation process, the appropriate classifier এটা "*eta*" is used in the corrected Assamese caption (C3) to express the proper meaning. Furthermore, in example Table 2(b), the word গজত "*gojot*" in translated Assamese caption (C2) is referring to the unit of length "Yard". This is a semantic error. In corresponding corrected Assamese caption (C3), this error is handled by replacing the incorrect word গজত with চোতালত "*sotalot*" (Lawn). Similarly, other semantic errors like কিক কৰিবলৈ "*kik koriboloi*" (Kick) and চোঁচৰা "*sosora*" (Shaggy) are handled by replacing with corresponding native Assamese words like লাথি মাৰিবলৈ "lathi mariboloi" and কেঁকোৰা "*kekora*" respectively.

Statistics of number of Images, Captions and Unique tokens present in the curated Flickr30K-AC and COCO-AC are shown in Table 1. The curated COCO-AC corpus contains *45,195* unique tokens consisting of *82.6%* nouns, *3.8%* verbs, *4.1%* adjectives and *9.5%* other parts-of-speech (see Table 3(a)). On the other hand, Flickr30K-AC contains *36,597* unique tokens out of which *71.5%* nouns, *6.7%* verbs, *8.7%* adjectives and *13.1%* other parts-of-speech (see Table 3(a)). Nouns, verbs and adjectives represent objects, activities and their attributes which are crucial to describe an image fluently. Further observations show that approximately *95%* of the captions in both Flickr30K-AC and COCO-AC dataset contains 4 to 24 tokens, with the distribution peaking around caption lengths of 10 and 11 (see Figure 2). This analysis can assist in setting the length of the generated captions.

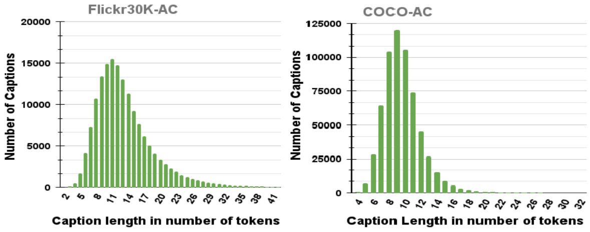


Figure 2: Flickr30K-AC and COCO-AC caption length distribution

Table 1: Statistics of Flickr30K-AC and COCO-AC dataset

Dataset	Images	Captions	Unique tokens
Flickr30K-AC	31783	158915	36579
COCO-AC	123287	616767	45195

4 Methodology

This section describes the proposed AIC model for Assamese caption generation that uses a Bi-LSTM decoder and bilinear attention.

The AIC uses a *Bi-LSTM* based decoder. It is a combination of two separate LSTM networks – a forward LSTM ($\mathbf{LSTM_F}$) and a backward LSTM ($\mathbf{LSTM_B}$) as shown in Figure 3(a). Generally, in AIC, a unidirectional LSTM predicts the current word w_t using visual features $\mathbf{V}$ of input image I and previous context information $\{w_1, \dots w_{t-1}\}$ by maximizing the probability $p\left(w_t|\mathbf{V}, \{w_1, \dots w_{t-1}\}\right)$. However, unidirectional LSTM only includes previous

Table 2: Examples from Flickr30K-AC and COCO-AC dataset with caption. C1 – Original English Caption, C2 – Translated Assamese Caption, C3 – Corrected Assamese Caption. Red colored words are translation error and blue colored words are correct words.

a		C1: A soccer player preparing to kick a soccer ball. C2: এজন ফুটবল খেলুৱৈয়ে এখন ফুটবল বল কিক কৰিবলৈ প্ৰস্তুতি চলাই আছে C3: এজন ফুটবল খেলুৱৈয়ে এটা ফুটবল লাথি মাৰিবলৈ প্ৰস্তুতি চলাই আছে
b		C1: Two young guys with shaggy hair look at their hands while hanging out in the yard C2: চোঁচৰা চুলি থকা দুজন ডেকা মানুহে গজত ওলাই থাকোঁতে তেওঁলোকৰ হাতলৈ চাওঁক C3: কেঁকোৰা চুলি থকা দুজন ডেকা মানুহে চোতালত ওলাই থাকোঁতে তেওঁলোকৰ হাতলৈ চাইছে

Table 3: Parts-of-Speech distribution for unique tokens present in Flickr30K-AC and COCO-AC.

POS Catagories	Number of Tokens with percentage	
	Flickr30K-AC	COCO-AC
Noun	27455 (75%)	39149 (86.2 %)
Verb	2562 (7%)	1792 (3.9%)
Adjective	3354 (9.2 %)	1946 (4.3%)
Others	3208 (8.2 %)	2555 (5.6 %)

context information and does not consider future context information $\{w_{t+1}, \ldots w_T\}$ while predicting w_t. The *Bi-LSTM* uses both previous and future context information with the help of $\mathbf{LSTM_F}$ and $\mathbf{LSTM_B}$ respectively. At decoding step t, $\mathbf{LSTM_F}$ computes the forward hidden state $\mathbf{h}_t^f = \mathbf{LSTM_F}\left(\mathbf{h}_{t-1}^f, \hat{\boldsymbol{v}}_t^f; \theta^f\right)$; while the backward hidden state $\mathbf{h}_t^b = \mathbf{LSTM_B}\left(\mathbf{h}_{t-1}^b, \hat{\boldsymbol{v}}_t^b; \theta^b\right)$ is computed with $\mathbf{LSTM_B}$ ($\mathbf{h}_t^f, \mathbf{h}_t^b \in \mathbb{R}^{d_h \times 1}\ \forall t$). Here, $\hat{\boldsymbol{v}}_t^f$ and $\hat{\boldsymbol{v}}_t^b$ are the attended visual features obtained from the bilinear attention modules (explain later) associated with $\mathbf{LSTM_F}$ and $\mathbf{LSTM_B}$ respectively. And, θ^f and θ^b are the respective parameters of $\mathbf{LSTM_F}$ and $\mathbf{LSTM_B}$. Next, two different word probability vectors $\mathbf{p}_t^f = SoftMax\left(W_p^f \mathbf{h}_t^f\right)$ and $\mathbf{p}_t^b = SoftMax\left(W_p^b \mathbf{h}_t^b\right)$ are respectively predicted by the forward and backward models. Here, $W_p^f, W_p^b \in \mathbb{R}^{l \times d_h}$ are linear transformations, and l is the word vocabulary size.

The word probability vector sequence $\mathbf{p}_t^f$ ($t = 1 \ldots T_f$) and $\mathbf{p}_t^b$ ($t = 1 \ldots T_b$) are subjected to beam search (Karpathy et al., 2014). This provides the respective captions generated by forward and backward models as $\tilde{S}_f = \{(\tilde{w}_t^f, \tilde{p}_t^f); t = 1 \ldots T_f\}$ and $\tilde{S}_b = \left\{(\tilde{w}_t^b, \tilde{p}_t^b); t = 1 \ldots T_b\right\}$. Here, $\tilde{w}_t^f, \tilde{w}_t^b$ are the most probable caption words with respective probabilities $\tilde{p}_t^f, \tilde{p}_t^b$ at decoding step t. The final caption $\tilde{S}$ is selected as follows (Cao et al., 2019).

$$\tilde{S} = \underset{r=f,b}{argmax}\left\{-\frac{\sum_{t=1}^{T_r} \tilde{p}_t^r \log(\tilde{p}_t^r)}{T_r}\right\} \tag{1}$$

The *bilinear attention* module (Figure 3(b)) transforms the visual features $\mathbf{V}$ to an attended embedding. It has two major components – first, an attention LSTM, and second an attention weight computation mechanism. Both $\mathbf{LSTM_F}$ and $\mathbf{LSTM_B}$ operate with similar attention LSTM networks $\mathbf{LSTM_A^f}$ and $\mathbf{LSTM_A^b}$ respectively. Here, $\mathbf{LSTM_A^f}$ is only discussed in the context of $\mathbf{LSTM_F}$ and the same description is also applicable for $\mathbf{LSTM_A^b}$ with respect to $\mathbf{LSTM_B}$. A pre-trained Faster-RCNN (Anderson et al., 2018) is used to identify the top-n regions of I. The d_v dimensional ResNet-101 embeddings of these regions are used as the visual features $\mathbf{V} = \{\boldsymbol{v}_1, \ldots \boldsymbol{v}_i, \ldots \boldsymbol{v}_n\}$ ($\boldsymbol{v}_i \in \mathbb{R}^{d_v \times 1}$). First, $\mathbf{LSTM_A^f}$ employs a top-down approach to generate a partially formed caption. The context information of the partially formed caption is captured in the hidden state $\mathbf{h}_t^{fA} \in \mathbb{R}^{d_h \times 1}$ of $\mathbf{LSTM_A^f}$. At each decoding step t, the input to $\mathbf{LSTM_A^f}$ includes the previous output of $\mathbf{LSTM_F}$, concatenated with the mean-pooled visual feature $\bar{\boldsymbol{v}} = \frac{\sum_{i=1}^n \boldsymbol{v}_i}{n}$.

$$\mathbf{h}_t^{fA} = \mathbf{LSTM_A^f}\left(\mathbf{x}_{t-1}^{fA}, \mathbf{h}_{t-1}^{fA}; \theta^{fA}\right), \mathbf{x}_{t-1}^{fA} = \left[\mathbf{h}_{t-1}^f : \bar{\boldsymbol{v}}\right] \tag{2}$$

Here, θ^{fA} is the parameter of $\mathbf{LSTM_A^f}$ and $[\ :\ ]$ denotes the concatenation operation.

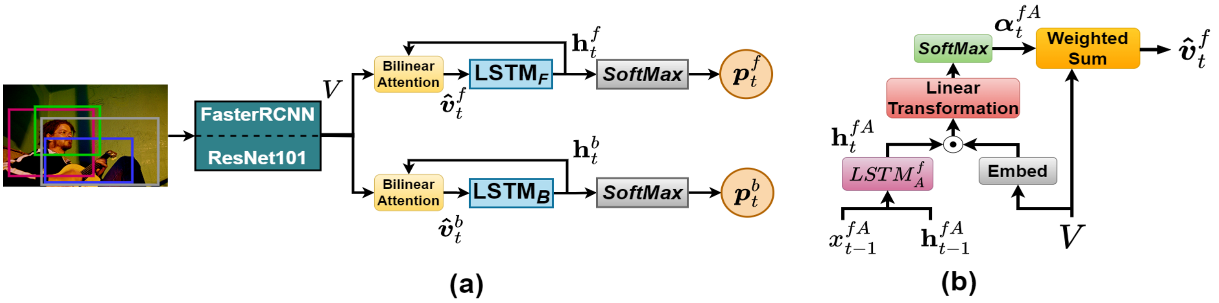


Figure 3: Functional block diagram of the proposed model. (a) Bi-LSTM model with separate LSTM networks for forward and backward processing. (b) Block diagram of the bilinear attention module.

Next, the attention weight computation mechanism generates a normalized attention weight $\alpha_t^{fA}[i]$ for each $\boldsymbol{v}_i$. The attention weights are computed using a low rank bilinear pooling operation involving $\mathbf{h}_t^{fA}$ and $\mathbf{V}$ in the following manner.

$$\alpha_t^{fA}[i] = SoftMax\left[\boldsymbol{\beta_{fA}}^T\left\{(W_a^{fh}\mathbf{h}_t^{fA}) \odot (W_a^{fv}\boldsymbol{v}_i)\right\}\ \right] \tag{3}$$

Here, $W_a^{fh} \in \mathbb{R}^{d_a \times d_h}$, $W_a^{fv} \in \mathbb{R}^{d_a \times d_v}$ and $\boldsymbol{\beta_{fA}} \in \mathbb{R}^{d_a \times 1}$ are linear transformations. Finally, the attended embedding $\hat{\boldsymbol{v}}_t^f \in \mathbb{R}^{d_v \times 1}$ for $\mathbf{LSTM_F}$ is obtained as

$$\hat{\boldsymbol{v}}_t^f = \sum_{i=1}^{n} \alpha_t^{fA}[i]\boldsymbol{v}_i \tag{4}$$

The model parameters θ^f, θ^b, W_p^f, W_p^b, θ^{fA}, W_a^{fh}, W_a^{fv}, $\boldsymbol{\beta_{fA}}$, θ^{bA}, W_a^{bh}, W_a^{bv}, $\boldsymbol{\beta_{bA}}$[2] are learned by minimizing the joint loss $\mathcal{L}_{Total}$ over the entire training dataset of input image and target caption pair as follows

$$\mathcal{L}_{Total} = \mathcal{L}_f(\tilde{S}_f, S_f) + \mathcal{L}_b(\tilde{S}_b, S_b) \tag{5}$$

Here, $\mathcal{L}_f$ and $\mathcal{L}_b$ are the respective losses associated with $\mathbf{LSTM_F}$ and $\mathbf{LSTM_B}$. The individual losses $\mathcal{L}_f$ and $\mathcal{L}_b$ are calculated at each decoding step using cross-entropy (Anderson et al., 2018). The same ground-truth caption is used in forward (S_f for $\mathcal{L}_f$) and backward (S_b for $\mathcal{L}_b$) order.

5 Experimental Setup

This section describes the model hyperparameters, dataset preparation, and baseline models used in this work

Model Hyperparameters – Pretrained Faster-RCNN (Anderson et al., 2018) is used to identify the top $N = 36$ image regions and their $d_v = 2048$ dimensional ResNet-101 embeddings are used as the visual features. The hidden state vector size of LSTM networks is set to $d_h = 1000$. The total loss ($\mathcal{L}_{Total}$) is minimized by the Adam optimizer with an initial learning rate of 0.0005. The batch size is kept as 32. The model is trained for 30 epochs.

Dataset Preparation – To perform experiments with the Flickr30K-AC and COCO-AC dataset, the Karpathy's split criteria (originally designed for English) (Karpathy et al., 2014) has been adopted. This splits the images of Flirck30K-AC as 29K for training, 1K for validation and testing, respectively. On the other hand, Karpathy's split for COCO-AC dataset is 113K images for training, 5K for testing and 5K for validation. Initially, basic preprocessing was performed by removing punctuation marks, followed by the creation of a vocabulary list. The vocabulary list contains words that occur more than five times in all captions of Flickr30K-AC and COCO-AC. This results in *8,534* and *12888* unique Assamese words in the vocabulary lists of Flickr30K-AC and COCO-AC, respectively. Additionally, captions containing more than 16 tokens are trimmed, allowing the maximum caption length to be of 16 tokens.

Baseline Models – This proposal is benchmarked against the following three baseline methods. First, the model proposed by Vinyals *et al.* (Vinyals et al., 2015) (Baseline-1). Here, the global image feature is provided as the initial hidden state input of the LSTM for caption generation. Second, the model proposed by Xu *et al.* (Xu et al., 2015) (Baseline-2) with ResNet101 (He et al., 2016) as encoder (instead of VGG16) and LSTM as de-

[2] Note that θ^{bA}, W_a^{bh}, W_a^{bv}, $\boldsymbol{\beta_{bA}}$ are the parameters of the attention module associated with $\mathbf{LSTM_B}$

Table 4: Performance comparison between proposed model and baseline for Flickr30K-AC.

Models	Decoder	Attention mechanism	Flickr30K-AC				
			B-1	B-2	B-3	B-4	C
Baseline-1	LSTM	-	51.1	32.5	21	13.3	31.9
Baseline-2	LSTM	Hard Attention	53.3	35.6	24.1	16.5	39.7
Baseline-3	LSTM	bottom-up	60.4	42.3	29.5	20.1	46.5
Proposed	LSTM	Bilinear	60.9	42.9	30	20.6	47.1
Proposed	Bi-LSTM	Bilinear	**61.5**	**43.4**	**30.4**	**21.1**	**48.3**

Table 5: Performance comparison between proposed model and baseline for COCO-AC.

Models	Decoder	Attention mechanism	COCO-AC				
			B-1	B-2	B-3	B-4	C
Baseline-1	LSTM	-	59.9	41.5	29.1	20.6	61.5
Baseline-2	LSTM	Hard Attention	62.1	45.1	31.8	23.2	68.3
Baseline-3	LSTM	bottom-up	67.1	49.6	36.5	27	80.8
Proposed	LSTM	Bilinear	67	49.7	36.6	27.1	81
Proposed	Bi-LSTM	Bilinear	**68.1**	**50.5**	**37.2**	**27.5**	**81.7**

coder. The Baseline-2 also incorporates attention mechanism to put more focus on salient regions of image based on previous context while generating caption. Third, the model proposed by Anderson *et al.* (Anderson et al., 2018) (Baseline-3). This model uses visual features of salient object regions instead of the whole image as bottom-up features with attention and LSTM as decoder.

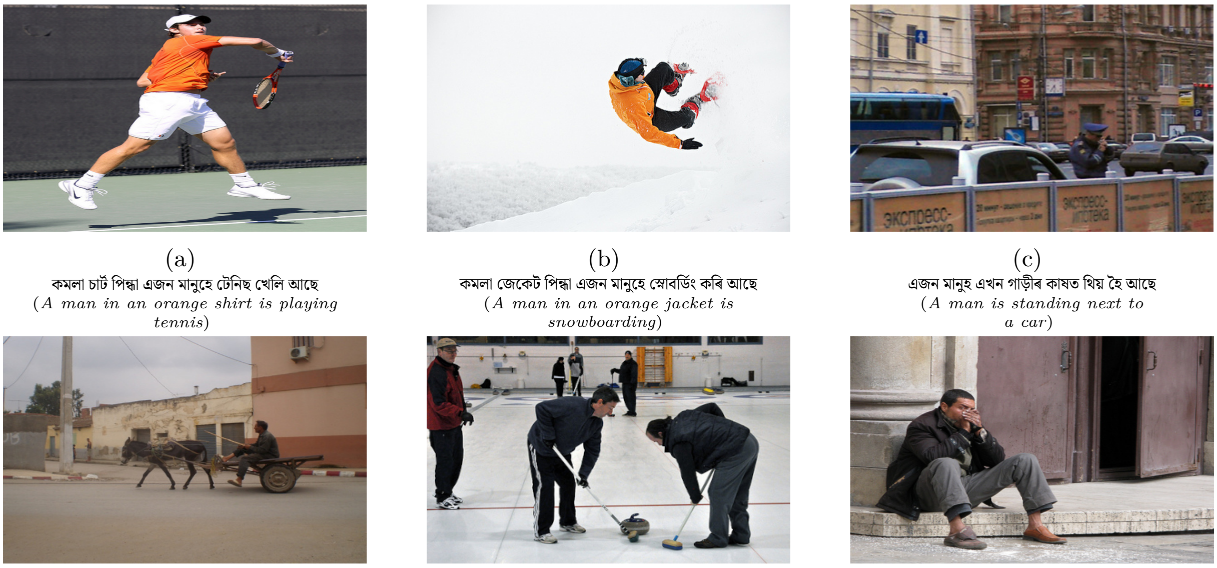
6 Results and Discussions

6.1 Quantitative Analysis

The proposed model is tested on 1000 and 5000 test images from the Flickr30K-AC and COCO-AC dataset respectively. The BLEU-n (*BLEU-1*: B-1, *BLEU-2*: B-2, *BLEU-3*: B-3, *BLEU-4*: B-4) and CIDEr (C) scores are computed for the generated captions and results are compared with baseline models. For comparison purpose the beam size value has been set to 3 for Flickr30K-AC and 6 for COCO-AC. The test results are shown in Table 5 and Table 4 for COCO-AC and Flickr30K-AC, respectively. Results presented in Table 4 and Table 5 indicate that bilinear attention combined with LSTM outperforms all other attention mechanisms achieving a CIDEr score of 47.1 and 81 for Flickr30K-AC and COCO-AC respectively. This can be attributed to the second order interactions between the visual features and context vector. Furthermore, the use of Bi-LSTM with bilinear attention yields the best results. For Flickr30K-AC it achieves a BLUE-4 score 21.1 and CIDEr score 48.3, while for COCO-AC it reaches BLUE-4 score 27.5 and CIDEr score 81.7. This can be attributed to the use of both forward and backward context information along with bilinear attention.

Additionally, Experiments were conducted to validate the impact of dataset curation process on model performance. For this experiment, the raw translated Assamese captions from Flickr30K (referred as Flickr30K-raw) and MSCOCO (referred as COCO-raw) were used without any correction of semantic and syntactic errors. The proposed Bi-LSTM with bilinear attention model is trained with Flickr30K-raw and COCO-raw. The results provided in Table 6 clearly demonstrate that the model trained on curated dataset significantly outperforms models trained on the raw dataset. This may be the semantic and syntactic uniformity present in Flickr30K-AC and COCO-AC datasets due to the curation process, thereby preserving the linguistic properties of the Assamese language.

Furthermore, captions obtained from the following systems are compared – (a) Assamese image captioning system, and (b) English image captioning system output translated to Assamese. For this, the proposed Bi-LSTM with



(a)
কমলা চাৰ্ট পিন্ধা এজন মানুহে টেনিছ খেলি আছে
(*A man in an orange shirt is playing tennis*)

(b)
কমলা জেকেট পিন্ধা এজন মানুহে স্নোবৰ্ডিং কৰি আছে
(*A man in an orange jacket is snowboarding*)

(c)
এজন মানুহ এখন গাড়ীৰ কাষত থিয় হৈ আছে
(*A man is standing next to a car*)

(d)
এজন মানুহে ৰাস্তাত ঘোঁৰা চলাই আছে
(*A man is riding a horse in the streets*)

(e)
দুজন মানুহে আইচ স্কেটিং কৰি আছে
(*Two people are ice skating*)

(f)
এজন মানুহে তেওঁৰ চেলফোনত কথা পাতি আছে
(*A man is talking on his cellphone*)

Figure 4: Examples of captions generated by the proposed model. First row shows accurate predictions and second row shows inaccurately generated captions.

Table 6: Performance of proposed Bi-LSTM and Bilinear attention on raw Assamese captions

Dataset	B-1	B-2	B-3	B-4	C
Flickr30K-Raw	59	41.1	28.5	18.1	45.1
COCO-Raw	66.9	48.3	34.4	24.5	75.8

bilinear attention model is trained with English captions of the MSCOCO dataset. Then generated English captions are translated to Assamese using the same Microsoft translator. Finally, the translated Assamese captions are compared with COCO-raw to maintain uniformity. Results from this experiment are shown in Table 7. This reveals that the model exclusively trained on Assamese dataset leads to a superior performance. This experiment demonstrates the importance of language specific image captioning system training for accurate and grammatically correct fluent caption prediction.

6.2 Qualitative Analysis

Figure 4 shows a few example captions generated for test images by the proposed model. In the first row of Figure 4, the generated captions successfully describe actual objects with their attributes and activity (e.g. Figure 4(a)). The use of bottom-up features enables the model to predict salient objects like "person", "car" even when there are other objects present in the image shown in Figure 4(c) . However, in the second row of Figure 4 the model mislabels the objects and associated activity. Consider the generated caption in example Figure 4(d). Here, the model incorrectly labels the object as a "horse" instead of a "horse-cart" as there are very few instances of "a man riding a horse-cart" in the dataset. In Figure 4(e) the model accurately identifies the objects, but mislabels the activities. This is due to the fact that the "white" background is predominantly associated with the activity "ice skating" in the dataset. Furthermore, the model occasionally detects objects that are not in the image. There are many images in the

Table 7: Performance of proposed model trained on MSCOCO English dataset and performance when Assamese captions are translated from intermediate English captions generated by the same model. This signifies the relevance of language specific training of image captioning models in Assamese language.

	B-1	B-2	B-3	B-4	C
Proposd model trained on MSCOCO English	77	60.7	46.5	35.2	112.5
Intermediate English Captions translated to Assamese	56.7	41.5	30.2	22.1	65.9

Table 8: Fluency and Adequacy rating scales and their meaning.

Fluency	**Adequacy**	**Rating**
Flawless	All information	4
Good	Most information	3
Not Fluent	Some information	2
Incomprehensible	No information	1

Table 9: Distribution of captions based on rating scale for both fleuncy and adequacy.

Rating Scale	Number of Captions	
	Fluency	Adequacy
1	30	37
2	93	131
3	211	195
4	166	137

dataset where a person's hand is near his head holding a "cellphone". Interestingly, a person with a similar posture in Figure 4(f) is described to have a "cellphone" which is absent in the image.

The quality of generated captions are evaluated through *fluency* and *adequacy* tests. The fluency determines the correctness of generated captions according to Assamese grammar rules. On the other hand, the adequacy describes how well the generated caption conveys visual information, including specifics like the quantity of objects and their characteristics present in the input image. For this test, five hundred (500) test images (from COCO-AC dataset) were randomly selected with their predicted captions. Two native Assamese speakers are selected as human annotators to rate the captions on a four point scale based on the rules described in Table 8. The results obtained from fluency and adequacy test is shown in Table 9. The results indicate that the model can generate fluent Assamese captions and can capture adequate visual information in the generated captions.

7 Conclusion

This work made two distinct contributions. First, the English captions of Flickr30K and MSCOCO dataset are translated to Assamese and manually corrected to form the Flickr30K-Assamese Caption dataset and COCO-Assamese Caption dataset. Second, a Bi-LSTM based decoder with bilinear attention is used for generating the captions in Assamese. The proposal is benchmarked against three baseline methods and is observed to provide competitive results in terms of BLEU-n and CIDEr scores. The proposed model can be extended to incorporate external knowledge and self-attention. This may enhance the diversity and relevance of the generated captions.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Shuang Bai and Shan An. 2018. A survey on automatic image caption generation. *Neurocomputing*, 311:291–304.

Pengfei Cao, Zhongyi Yang, Liang Sun, Yanchun Liang, Mary Qu Yang, and Renchu Guan. 2019. Image captioning with bidirectional semantic attention-based guiding of long short-term memory. *Neural processing letters*, 50:103–119.

C Chandramouli and Registrar General. 2011. Census of india. *Rural Urban Distribution of Population, Provisional Population Total. New Delhi: Office of the Registrar General and Census Commissioner, India.*

Ringki Das and Thoudam Doren Singh. 2022. Assamese news image caption generation using attention mechanism. *Multimedia Tools and Applications*, 81(7):10051–10069.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 15–29. Springer.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Andreas Holzinger, Anna Saranti, and Heimo Mueller. 2021. Kandinskypatterns–an experimental exploration environment for pattern analysis and machine intelligence. *arXiv preprint arXiv:2103.00519.*

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4634–4643.

Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. *Advances in neural information processing systems*, 27.

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2891–2903.

Siming Li, Girish Kulkarni, Tamara Berg, Alexander Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228.

Santosh Kumar Mishra, Rijul Dhir, Sriparna Saha, and Pushpak Bhattacharyya. 2021a. A hindi image caption generation framework using deep learning. *Transactions on Asian and Low-Resource Language Information Processing*, 20(2):1–19.

Santosh Kumar Mishra, Rijul Dhir, Sriparna Saha, Pushpak Bhattacharyya, and Amit Kumar Singh. 2021b. Image captioning in hindi language using transformer networks. *Computers & Electrical Engineering*, 92:107114.

Santosh Kumar Mishra, Sushant Sinha, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A deep learning based framework for image paragraph generation in hindi. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 792–800.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alexander Berg, Tamara Berg, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756.

Prachurya Nath, Prottay Kumar Adhikary, Pankaj Dadure, Partha Pakray, Riyanka Manna, and Sivaji Bandyopadhyay. 2022. Image caption generation for low-resource assamese language. In *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, pages 263–272.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10971–10980.

Dhrubajyoti Pathak, Sukumar Nandi, and Priyankoo Sarmah. 2022. Aspos: Assamese part of speech tagger using deep learning approach. In *2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–8. IEEE.

Matiur Rahman, Nabeel Mohammed, Nafees Mansoor, and Sifat Momen. 2019. Chittron: An automatic bangla image captioning system. *Procedia Computer Science*, 154:636–642.

K Samudravijaya. 2021. Indian language speech label (ilsl): A de facto national standard. In *Advances in Speech and Music Technology: Proceedings of FRSM 2020*, pages 449–460. Springer.

Faisal Muhammad Shah, Mayeesha Humaira, Md Abidur Rahman Khan Jim, Amit Saha Ami, and Shimul Paul. 2021. Bornon: Bengali image captioning with transformer-based deep learning approach. *arXiv preprint arXiv:2109.05218*.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2022. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):539–559.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. 2016. Image captioning with deep bidirectional lstms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 988–997.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.