SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Region-Based Attention With Hierarchical Bilstm For Image Captioning

Chaitanya S. Bhosale*[1] , Dr. Pradip Salve[2] , Dr. Vishal Shirsath[3]

[1]Ajinkya DY Patil University, Pune, Maharashtra, India chaitanya.bhosale@adypu.edu.in
[2]Ajinkya DY Patil University, Pune, Maharashtra, India pradip.salve@adypu.edu.in
[3]Ajinkya DY Patil University, Pune, Maharashtra, India vishal.shirsath@adypu.edu.in
*Corresponding author

ABSTRACT
Image captioning requires expressing complex visual scenes in coherent, contextually relevant natural language, but traditional encoder-decoder frameworks rely on a single global feature and a unidirectional decoder, limiting fine-grained understanding. This study proposes a framework integrating region-based feature extraction with hierarchical bidirectional decoding to generate semantically rich captions, aimed at helping visually impaired individuals perceive their surroundings without relying solely on touch. A Region-based Convolutional Neural Network (RCNN) extracts object-level visual features through region proposals, refined using a visual attention mechanism that focuses on the most relevant regions during generation. Refined features are decoded using a two-level Hierarchical Bidirectional LSTM (H-Bi-LSTM), where the lower level models word-level linguistic dependencies and the higher level captures sentence-level semantics. The key novelty lies in this joint region-based attention and hierarchical bidirectional decoding, which explicitly separates word-level syntax from sentence-level semantics. Generated captions are further converted into audible speech in English, Hindi, and Marathi through a text-to-speech module, enabling multilingual, accessibility-oriented delivery. Evaluated on MSCOCO and Flickr30K, the model achieves BLEU@4 scores of 39.3 and 38.2, with corresponding improvements in METEOR and CIDEr over existing methods.

Keywords-Image captioning; Natural Language Processing; Region Based Convolutional Neural Networks (RCNN); Hierarchical Bidirectional Long Short-Term Memory (H-Bi-LSTM); Multilingual Captioning; Text-to-Speech; Assistive Technology for the Visually Impaired.

INTRODUCTION

The process of image captioning is to generate descriptive textual content for inputted image which is a challenging process, especially if image contains multiple objects and complex scenery. Finding and identifying everyday objects in the surrounding environment is an enormous undertaking for individuals who are blind or visually impaired, who are often dependent on other people or on their sense of touch and smell, which can be dangerous and unreliable in several situations. Because of its potential applications image captioning has gained significant attention. Image captioning can have multiple use cases such as aiding visually impaired individuals, image classification based on their content and human-computer interaction [1-3]. A comprehensive review of image captioning techniques, their deep learning advancements, and associated limitations has also been presented in the literature [38]. Automatic annotation techniques have likewise been explored in other application domains, such as annotating query results retrieved from deep web databases, underscoring the broader relevance of annotation-driven approaches beyond image captioning [36]. To complete image captioning requires two main tasks, first is to understand visual content of an image by extracting the features that gives meaning to an image and secondly express that understanding in meaningful and grammatically correct textual description. Traditional

image captioning techniques were dependent on rule based language models and pre-defined features which often failed to generate accurate captions and were limited in their descriptive capabilities. With the advancement in deep learning, hybrid approaches have emerged, integrating Convolution Neural Network (CNN) for visual feature extraction and Recurrent Neural Networks (RNN) for language generation [4-7]. These type of hybrid models showed significant enhancement in capturing complex relationships between objects in an image and describing them in natural language. CNNs became a default standard for feature extraction because of their ability to learn hierarchical representations of images. They can capture spatial and semantic information efficiently which makes them suitable for understanding the content of an image but they cannot understand temporal dependencies needed to generate contextual sentence [8].

To address aforementioned problem, RNNs, particularly LSTM networks, have been widely adopted for sequence modelling in image captioning [32], [33]. LSTMs are able to capture long-range dependencies in sequential data, which is crucial for generating grammatically sound and contextually relevant captions [9-11]. Furthermore, the use of bidirectional LSTMs (Bi-LSTMs) enhances ability of the model to capture contextual information by processing sequences in both forward and backward directions. This bidirectional approach make sure that the generated captions consider the complete context of the input features, resulting in more coherent and descriptive outputs. More recently, self-attention based Transformer architectures [34] have also been explored for sequence modelling, motivating the attention-driven decoding strategy adopted in this work.

CNN-RNN encoder decoder models are solely depend on a single global image feature representation resulting in the loss of fine-grained object-level semantics and minimal understanding of complex scenes containing multiple entities. Lack of region-level feature extraction limits models ability to link visual regions with respective linguistic tokens resulting in generic and incomplete captions. Most decoders are unidirectional such as LSTM which generates captions in sequence from left to right which are unable to capture long range dependencies resulting in limited understanding of contextual information. These baseline decoders lack hierarchical language modelling which treats sentence semantics and word syntax uniformly. Graph-structured context modelling has also been explored as an alternative means of capturing object relationships during decoding [35], though such approaches add structural complexity that depends heavily on the quality of the underlying scene graph. To address these limitations a framework is designed that combines RCNNs for visual feature extraction, structured visual attention and a Hierarchical Bidirectional LSTM (Bi-LSTM) network for sentence generation. The RCNN component extracts rich semantic features from input images, which are then fed into the Deep Bi-LSTM to generate captions that are both syntactically correct and semantically meaningful. By leveraging the strengths of RCNNs and Bi-LSTMs, our approach effectively bridges the gap between visual and textual representations.

We have proposed a novel Hierarchical BiLSTM for multilevel semantic understanding and performance of the proposed method evaluated on two benchmark datasets: MSCOCO and Flickr30K. The MSCOCO dataset provides a large and diverse set of images paired with multiple human-annotated captions, making it ideal for training robust models. Flickr30K, on the other hand, contains domain-specific images with captions that emphasize diverse linguistic styles and challenges. These datasets enable comprehensive evaluation and comparison of our model's generalizability and effectiveness in real-world scenarios. Evaluated experimental results shows that the proposed model is able to generate competitive results, surpassing or matching existing state-of-the-art methods on standard evaluation metrics such as BLEU, METEOR, and CIDEr. Additionally, we presented qualitative analyses to illustrate ability of the model to generate precise and contextually rich captions across diverse image types. The remaining of this paper is structured as follows: Section 2 discusses details the architecture of our proposed RCNN and Deep H-Bi-LSTM-based framework. Section 3 includes experimental setup, datasets details, performance evaluation in standard evaluation metrics. Section 4 concludes the paper with a summary of findings and potential improvements.

MATERIALS AND METHODS

As illustrated in Figure 1, the proposed architecture consists of three core components:

First is region-based visual feature extraction using Region-based Convolutional Neural Networks (RCNN), second is attention-based visual feature refinement and third is sentence decoding using hierarchical Bi-LSTM. First the input image is processed to detect important regions and extract object level features using RCNN. This step preserves spatial and object-centric information instead of single global image feature.

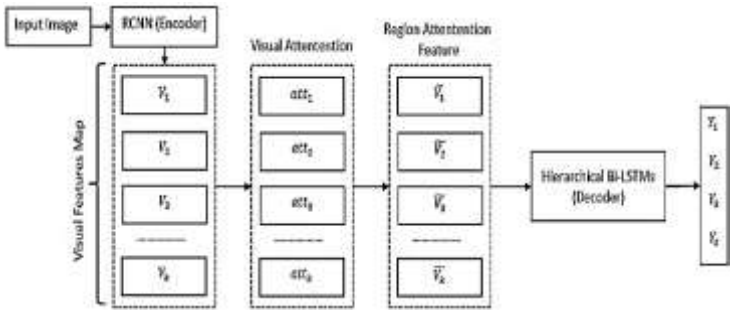


Fig. 1. Basic architecture of proposed caption generation model

In second step extracted region features are weighted dynamically to highlight the importance of each region at different time stamp using visual attention mechanism. This attention-guided mechanism allows model to pick most important visual regions during caption generation resulting in improved visual–linguistic alignment. Finally attended visual features are processed by two-level hierarchical bidirectional long short term memory (H-Bi-LSTM) decoder. Here the lower layer models word-level linguistic dependencies while the higher layer captures sentence-level semantics. This integrated design guarantees clear separation between visual representations, attention guided feature selection and hierarchical language modelling thereby addressing the limitations of traditional encoder-decoder image captioning models.

RCNN Encoder

The image is processed using a pre-trained VGG-16 based Region-based Convolutional Neural Network (RCNN) to extract a visual features map. Convolutional layers extract low- to high-level spatial patterns at different scales, pooling layers downsample the resulting feature maps while retaining salient information, and fully connected layers integrate these features for high-level abstraction; the output layer then performs object detection using bounding-box regression and class-probability estimation to localise and identify candidate objects within the image. There are several modern object detection models such as Faster-RCNN, Mask-RCNN, or transformer-based detectors like DETR. We have integrated RCNN because of its architectural simplicity and modularity. RCNN focuses on object centric regions that contains fine-grained semantic information which is important for accurate caption generation. The main aim of proposed work is to evaluate the effectiveness of hierarchical language modelling when integrated with region-level visual semantic. Inference speed is not the main focus of the study so the computation overhead of RCNN does not affect the result in any way. In future RCNN can be replaced by more advanced visual encoder like Faster R-CNN, Mask R-CNN, or transformer-based models such as data-efficient image transformers, which have shown strong transfer learning performance in other visual recognition tasks [28], to enhance computational efficiency and scalability. CNN-based backbones have similarly proven effective for real-time recognition tasks such as traffic sign recognition, demonstrating the versatility of CNN architectures across diverse object recognition applications [37]. Notation description is given in Table I.

I. NOTATION DESCRIPTION FOR MODEL ARCHITECTURE

Notation	Definitions
I	Input Image
V	Feature vector
att	Attention weights
R	Aggregated region feature
y_{t-1}	Previous word embedding
h_t	Hidden state
a_t, b_t	Hidden states of LSTM units at time step t in forward and backward directions, respectively
X^a, X^b	Input data flowing into the forward and backward layers
Y^a, Y^b, Y^o	Output data at various stages in the hierarchy
I_t	Input at time step t, representing visual features from an image.
O_t	Final output at time step t, combining results of forward and backward layers

The output of RCNN is a set of feature vectors $\{V_1, V_2, V_3 \dots \dots V_k\}$ where each vector $V_i \in R^d$ represents the feature embedding for a region in the image. Mathematically, the transformation is: $\{V_1, V_2, V_3 \dots \dots V_k\} = RCNN(I)$ Here, k is the number of regions and d is the dimensionality of each feature vector.

Visual Attention Module

This module applies attention weights over the visual features from the RCNN. Attention weights $\{att_1, att_2, att_3, \dots \dots att_k\}$ are calculated to focus on the most important regions. The attention mechanism computes:

$$att_i = \frac{exp(e_i)}{\sum_{j=1}^{k} exp(e_j)} \quad (1)$$

$$e_i = f(V_i, h_{t-1}) \quad (2)$$

e_i is the relevance score for region i, based on the feature V_i and the previous hidden state h_{t-1}. f is a scoring function, often implemented as a neural network. The attended features are:

$$\widetilde{V_i} = att_i.V_i \quad (3)$$

Region Attention Feature Module

After applying the visual attention, a Region Attention Feature is computed by refining and aggregating the features:

R_j=Aggregate $(\{\widetilde{V_i}\})$ (4)

Here, R_j represents the refined feature vector for a specific region of interest and the aggregation function is be a weighted sum of features.

Here we have used Bahdanau-style additive soft attention mechanism in order to align visual region features with the decoding process. This attention mechanism evaluates relevance score using learned scoring function that combines visual feature of each region and the previous hidden state of the decoder.

Hierarchical Bi-LSTMs Decoder

Figure 2. represents a hierarchical Bi-LSTM (H-Bi-LSTM) architecture with multiple forward and backward layers for sequential data processing, which is applied to image captioning.

In the proposed approach, the hierarchical Bi-LSTM is used for caption generation from the obtained feature vector provided by RCNN. As shown in Figure 2. Architecture consist of two layer of Bi-LSTMs that are connected in forward and backword directions hence the name hierarchical Bi-LSTM. Stacking of two Bi-LSTMs enables system to not only captures lower-level linguistic patterns but also high-level semantics, that is first layer captures linguistic patterns and output of first layer is given as input to second layer to generate semantically rich captions.

Lower–level Bi-LSTM processes caption sequentially at each time step where the input contains embedding of the previously generated word as well as visual feature vector extracted by RCNN with attention weights. Lower–level Bi-LSTM captures short term dependencies producing a sequence of word-level hidden states. Higher level Bi-LSTM processes aggregated representation of the lower-level hidden states. Mainly those hidden states that are generated by lower level Bi-LSTM at each time step which is summarised using pooling operation to generate compact representation that encodes the evolving sentence context. Higher level Bi-LSTM then generate a sentence-level context vector by processing summarized representation which captures global semantic information like scenes and object relationships. Generated sentence level context vector is then concatenated with word level hidden state and passed to output layer for word prediction.

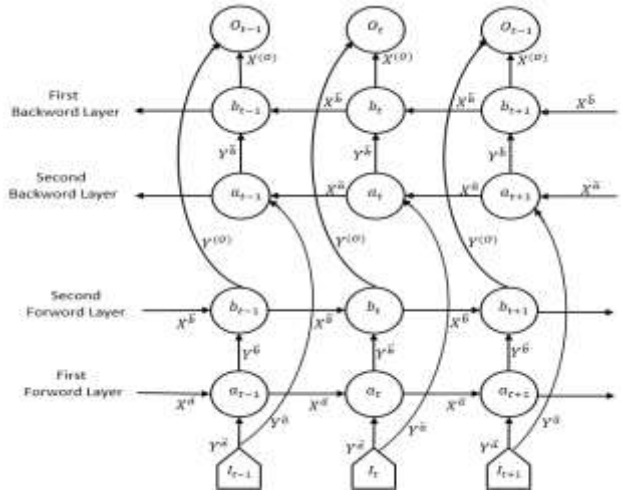


Fig. 2. Information flow within Hierarchical Bi-LSTMs.

Proposed model does not use future ground-truth tokens during inference. Caption is generated in left to right manner while the bidirectionality is performed in hidden-states at each decoding step instead of output token sequence.

This module generates the final caption sequence $\{Y_1, Y_2, Y_3, \dots \dots Y_t,\}$ based on the features. Hierarchical Bi-LSTM has two levels first is sentence-level LSTM which controls the generation of sentences and second is Word-level LSTM which decodes features into words.

For each time step t:

$$h_t = LSTM([R, y_{t-1}], h_{t-1}) \quad (5)$$

The output word probabilities are:

$$P(y_t|y < t, I) = softmax(Wh_t + b) \quad (6)$$

Where, W and b are learned parameters.

As shown in Figure 2. The input sequence $\{I_1, I_2, \dots \dots I_t\}$ for time t, fed into the network which is then processed in two directions.

Forward Layers (First and Second)

These process the input I^t sequentially from left to right.

The input enters hidden state a_t in forward direction that is $\{\overrightarrow{a_{t-1}}, \overrightarrow{a_t}, \overrightarrow{a_{t+1}}\}$ time step t is updated using:

$$a_t = LSTM_f(X_t^a, a_{t-1}) \quad (7)$$

Where, X_t^a: Input feature for the forward layer at time t. a_{t-1}: Hidden state from the previous forward time step. Outputs from forward layers are denoted as Y^a and passed to the higher hierarchical layers.

Backward Layers (First and Second)

This step process the input sequence in reverse order, from right to left**.** The hidden state b_t at time step t is updated as:

$$b_t = LSTM_b(X_t^b, b_{t+1}) \quad (8)$$

Where, X_t^b: Input for the backward layer at time t. b_{t+1}: Hidden state from the next time step (in reverse order). Outputs from backward layers are denoted as Y^b and also passed to higher layers.

Outputs at Each Hierarchical Level

The outputs from both forward and backward layers are combined to generate a context-aware representation at each time step:

$$Y_t^{(O)} = f(Y_t^a, Y_t^b) \quad (9)$$

f: A concatenation function that merges the outputs of forward and backward passes. These combined outputs $Y_t^{(O)}$. The combined output is used to predict words in the caption sequentially at each time step t.

Text-to-Speech Module

To make the generated captions directly usable by visually impaired individuals, the caption produced by the H-Bi-LSTM decoder is passed to a text-to-speech (TTS) module that converts the predicted text into audible speech output. In addition to English, the caption text is translated and synthesised into Hindi and Marathi, allowing the same visual description to be delivered as narrated audio in multiple regional languages. This multilingual audio pathway removes the need for the user

to read on-screen text and lets the system describe the surrounding scene aloud in a language the listener is comfortable with, directly supporting the accessibility objective motivating this work.

RESULTS AND DISCUSSION

Proposed Hierarchical BiLSTM model is trained and tested on two standard datasets MSCOCO [12] and Flickr30K [13]. We have performed extensive experiments and tested the model on standard metrics while comparing it with other existing models. Proposed framework is implemented using pre-trained VGG-16 based RCNN to extract region-level visual features, and each region is represented by a 2048-dimensional feature vector while the dataset split is 80:20 from which captions are tokenized and converted into word indices using a fixed vocabulary generated from the training split and each word is embedded into 256 dimensional feature vector. Decoder consist of two-level Hierarchical Bidirectional LSTM and each layer contains 512 hidden units. Framework employed soft visual attention mechanism over extracted region features by RCNN where attention weights are evaluated using differentiable scoring function and normalized using softmax function. Model is trained with learning rate of 0.0001 and batch size of 64. Adam optimizer is used in training with a maximum of 30 epoch with early stoppage based on validation loss.

Datasets

To assess proposed work we utilized the MSCOCO [12] dataset with over 330,000 images, which is a benchmark dataset containing comprehensive collection of images sourced from various websites with diverse subjects, including people, animals, buildings, parks vehicles, and other key objects. Each image is paired with five captions which are different form each other. These captions offer detailed description about images which can be used as ground truth reference.

Flickr30K is an extended version of the Flickr8K dataset containing 31,783 images, each image is annotated with five descriptive captions. This dataset is mainly used in image captioning tasks. Following the publicly accessible splits provided by [13], the dataset is divided into 29,000 images for training, 1,000 for validation, and 1,000 for testing.

Training and Validation Analysis

Fig. 3 shows the training and validation loss of the Bi-LSTM decoder over 12 epochs. At the start of training, the training loss drops from 6.82 to 4.17, indicating that the model is learning meaningful mappings from image features to captions, while the validation loss decreases from 6.99 to 6.37 over the same early epochs, suggesting good generalization at the start of training. During epochs 6 to 12, training loss keeps falling to about 2.1, showing that the model continues to fit the training data well; however, the validation loss starts fluctuating between roughly 6.7 and 7.4 over the same epochs, which is a sign of overfitting, that is, the model memorises the training data but struggles to generalize equally well to unseen validation samples. This trend motivated the early-stopping criterion based on validation loss used during training.

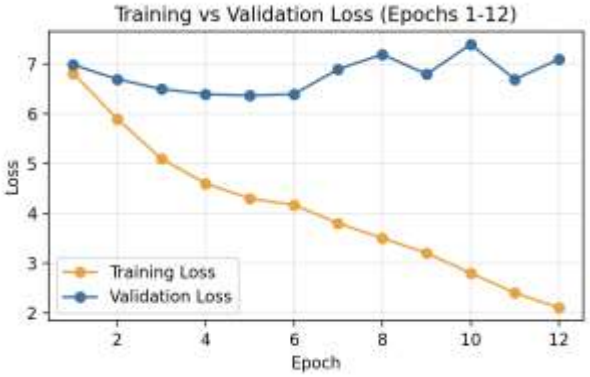


Fig. 3. Training and validation loss over 12 epochs for the Bi-LSTM decoder.

Qualitative Analysis

To assess proposed work we utilized the MSCOCO and Flickr30K datasets. Table II shows the quantitative comparison of captions generated by various models and ground truth which shows how well models elaborate content in an image. In first image ground truth is “two men that are standing in a kitchen”. Caption generated by Bi-LSTM [1] is “two men standing in a kitchen” which is quite close to ground truth and very simple sentence construction. Caption generated by proposed H-Bi-LSTM is “two men standing in a kitchen in front of a stove” which is mode detailed and shows semantic and spatial understanding. In second image proposed H-Bi-LSTM managed to capture details like “packed,” “sitting on top,” and “overloaded,” which is very close to ground truth. In third image proposed model identifies color incorrectly that is “blue” instead of gray. Evaluated results on MSCOCO dataset suggests that the proposed model generates context aware, semantically rich and descriptive captions that align with the ground truth. Table III shows comparative analysis of various

image captioning models with standard metrics tested on MSCOCO dataset. Among all the models proposed model achieves balanced and competitive results across all metrics. Existing models like S-Att Bi-LSTMs and WCNN with VAPN perform better than rest of the models especially in CIDEr and BLEU-4 but fall short in METEOR showing that when it comes to capturing semantic details these models fell short while proposed model shows effective results across all metrics, generating semantically accurate captions. Table II shows the caption generated by BiLSTM and H-BiLSTM . Caption generated by BiLSTM is “a dog is running through a field" which offers limited scene context while the proposed models adds visual attributes like “black and white dog”, “white fence” and “grass” giving more detailed description of the scene. In second caption proposed model adds “table”, “umbrella” and café which reflect the environment in the scene capturing meaningful relationship. Table IV presents a quantitative comparative analysis of various captioning models across standard metrics. AIC-SSAIDL [14] archives strong BLEU score as well as high METEOR score but lower CIDEr indicating overfitting or lack of diversity. DenseAtt [15] archives highest CIDEr score suggesting alignment with human captions but has low METEOR and ROUGE-L score. Bi-LSTM+VGGNet [16] achieves average scores across all metrics but lack detailed semantic modelling.

Evaluated results shows that the proposed H-Bi-LSTM model achieved BLEU@4 score of 39.3 on MSCOCO dataset while the transformer based captioning models achieved BLEU@4 scores of 40 that is because of large-scale pre-training and self-attention mechanisms. Results achieved by proposed model is competitive within the class of recurrent region based captioning models. As compared to conventional encoder–decoder models proposed H-Bi-LSTM model achieves competitive results across all metrics suggesting model maintained balance between syntactic correctness (BLEU), semantic details (METEOR) as well as relevance (CIDEr).

II. COMPARATIVE QUALITATIVE ANALYSIS OF EXISTING MODELS AND PROPOSED MODEL WITH GENERATED CAPTIONS ON MSCOCO AND FLICKR30K DATASET.

Image from MSCOCO	Captions	Image from Flickr30K	Captions
	**BiLSTM [1]**: two men standing in a kitchen **Proposed H-BiLSTM**: two men standing in a kitchen **in front of a stove**		**BiLSTM [17]**: a dog is running through a field **H-BiLSTM**: a **black and white** dog running through a **grass** near a **white fence**
	**BiLSTM [1]**: a group of people riding on the back of a truck **Proposed H-BiLSTM**: a **packed group** of men **sitting on top** of a overloaded pickup truck		**BiLSTM [17]**: a group of people sitting in front of a store **H-BiLSTM**: a group of people sitting **at a table under umbrella outside a cafe**
	**BiLSTM [1]**: a bike standing in front of a wall **Proposed H-BiLSTM**: a gray bicycle locked to a **blue** metal doors		**BiLSTM [17]**: a young girl is eating a meal **H-BiLSTM**: a **toddler** eating a meal from a **glass bowl**

III. COMPARATIVE ANALYSIS OF CURRENT EXISTING MODELS WITH PROPOSED H-BILSTM MODEL BASED ON VARIOUS MATRICES EVALUATED ON MSCOCO DATASET. PLACEHOLDERS ("-") HAVE BEEN USED WHERE RESULTS WERE NOT EXPLICITLY AVAILABLE

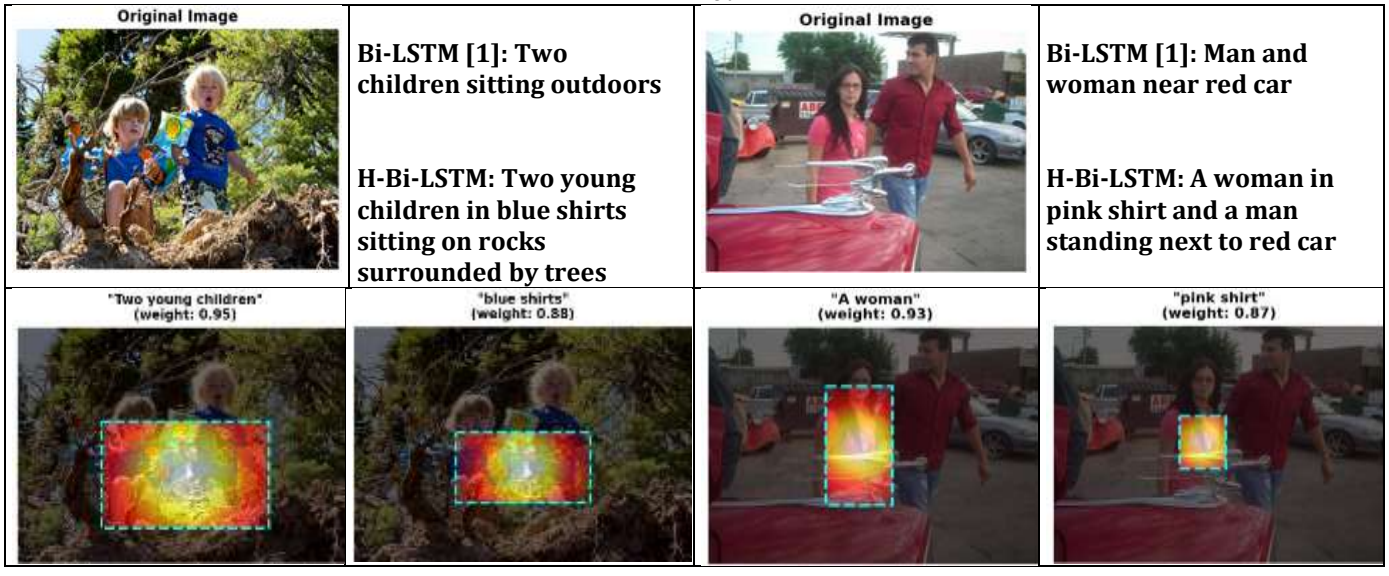
Models	B@1	B@2	B@3	B@4	METEOR	ROUGE-L	CIDEr
S-Att Bi-LSTMs [1]	79.3	64.0	49.1	37.5	28.9	58.1	121.3
Semantic Guided Attention [3]	78.6	-	-	36.0	27.6	57.7	120.9
Scene Graph with GA and CG [18]	-	**78.1**	-	**56.7**	-	**70.2**	57.8
WCNN with VAPN and LSTM [19]	78.5	62.0	49.1	38.2	28.9	58.3	124.2
CSF [20]	76.4	60.2	46.1	35.0	27.1	56.1	110.2
POS with BiLSTM [21]	75.7	-	-	33.2	30.1	60.3	97.6

Multi-Layer Attention [22]	70.50	56.62	49.60	33.07	**51.77**	-	-
AIC-SSAIDL [14]	**80.40**	62.94	47.81	38.04	33.58	-	**137.45**
MGAN [23]	80.06	65.2	50.7	38.5	28.9	58.6	129.6
DenseAtt [15]	-	-	51.0	34.0	24.8	56.5	118.3
HAF [24]	75.9	59.5	45.4	34.4	26.8	-	109.0
HAD [25]	80.0	64.6	50.1	38.2	28.5	58.1	123.3
Stack-VS Attention [26]	79.0	63.4	48.9	37.2	27.8	57.5	118.9
VGGNet NeuralTalk2-T-oe(stage3) [27]	74.0	56.7	44.3	31.3	25.5	53.4	98.3
TDA+GLD [13]	79.0	63.0	48.2	36.3	27.7	57.1	117.9
Proposed	80.3	63.4	**50.5**	39.3	39.8	58.6	126.2

IV. COMPARATIVE ANALYSIS OF CURRENT EXISTING MODELS WITH PROPOSED H-BILSTM MODEL BASED ON VARIOUS MATRICES EVALUATED ON FLICKR30K DATASET. PLACEHOLDERS ("-") HAVE BEEN USED WHERE RESULTS WERE NOT EXPLICITLY AVAILABLE

Models	B@1	B@2	B@3	B@4	METEOR	ROUGE-L	CIDEr
WCNN+VAPN+LSTM [19]	70.1	49.4	35.8	27.2	21.7	-	67.3
LSTM [21]	59.0	-	-	18.5	23.5	49.9	30.8
BiLSTM [21]	64.8	-	-	22.1	26.3	54.4	50.4
Bi-F-LSTM [17]	59.9	41.2	27.3	18.1	-	-	-
AIC-SSAIDL [14]	71.42	**62.56**	53.81	41.18	35.09	-	61.89
DenseAtt [15]	-	-	52.0	39.1	26.0	51.0	**209.1**
Bi-LSTM+VGGNet [16]	62.1	42.6	28.1	19.3	-	-	-
Proposed H-BiLSTM	**78.2**	61.4	**48.9**	38.2	33.8	**55.7**	123.6

V. HEATMAP FOR GENRATED CAPTIONS BY H-BI-LSTM EVALUATED USING FLICKR30K DATASET



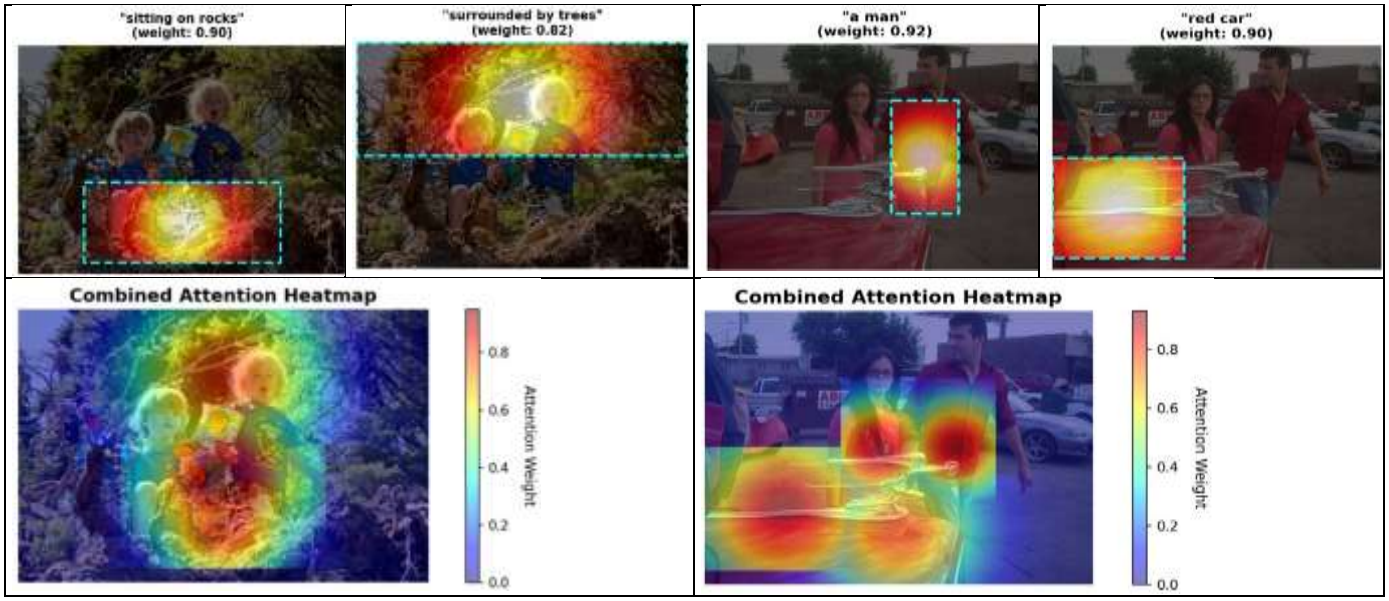


Table V shows heatmap of generated captions. Traditional BiLSTM generated caption is "Two children sitting outdoors" where H-BiLSTM generated caption is "Two young children in blue shirts sitting on rocks surrounded by trees", showing clear difference between generated captions by two decoders.

Caption generated by traditional Bi-LSTM clearly describes image in detail such as it has added age attribute ("young"), clothing detail ("blue shirts"), specific surface ("on rocks"), environmental context ("surrounded by trees") suggesting that proposed methods has better spatial understanding.

Multilingual Caption Generation

Beyond English, the generated captions were passed through the text-to-speech module to produce spoken output in Hindi and Marathi. For an image of two construction workers taking a break, the proposed H-Bi-LSTM predicted the caption "two construction workers sit on base taking break," which was then rendered as audio in English, Hindi, and Marathi, giving the same description in each language. For an image of a couple seated on grass with a baby stroller, the model generated "couple sit on the grass with baby and stroller" and again produced corresponding Hindi and Marathi audio narration. These examples indicate that the captioning pipeline generalises beyond English generation and can deliver the same scene description audibly in multiple regional languages, which is central to the accessibility goal motivating this work.

CONCLUSION

Presented paper proposes an image captioning framework integrating Region-based Convolutional Neural Networks (RCNN) for feature extraction along with an attention mechanism which focuses on important regions within an image. A novel two-level Hierarchical Bidirectional LSTM (H-Bi-LSTM) decoder to generate semantically and contextually rich captions. Proposed hierarchical design allows model to capture word-level linguistic dependencies and semantics on a sentence level. Generated captions are further converted into audible speech in English, Hindi, and Marathi through an integrated text-to-speech module, enabling the framework to assist visually impaired individuals in independently understanding their surroundings across languages. Model is evaluated on MSCOCO as well as Flickr30k dataset. Evaluated results demonstrate that proposed model outperforms baseline Bi-LSTM architectures across all standard metrics. Qualitative results shows that the model can generate detailed human like descriptive captions. Proposed work can be extended with additional Indian and international languages, using multilingual transformers like mBERT or mT5. Future extensions may also explore large-scale vision-language pretraining and instruction-tuned multimodal architectures [29]-[31], which have demonstrated strong few-shot and zero-shot captioning capabilities.

CONFLICT OF 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.

ACKNOWLEDGMENT

I want to thank my guide, Dr. Pradip Salve and Dr. Vishal Shirsath their mentorship throughout the paper writing process. their deep knowledge, expertise, and constant help have greatly influenced writing of this paper. I deeply appreciate their guidance.

REFERENCES

[1] Z. Han, C. Ma, Z. Jiang and J. Lian, "Image caption generation using contextual information fusion with Bi-LSTM-s," in IEEE Access, vol. 11, pp. 134–143, 2023, doi: 10.1109/ACCESS.2022.3232508.

[2] Y. Dongare, B. M. Hardas, R. Srinivasan, V. Meshram, M. G. Aush and A. Kulkarni, "Deep neural networks for automated image captioning to improve accessibility for visually impaired users," International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 2s, pp. 267–281, 2023.

[3] D. A. Hafeth, S. Kollias and M. Ghafoor, "Semantic representations with attention networks for boosting image captioning," in IEEE Access, vol. 11, pp. 40230–40239, 2023, doi: 10.1109/ACCESS.2023.3268744.

[4] A. Ueda, W. Yang and K. Sugiura, "Switching text-based image encoders for captioning images with text," in IEEE Access, vol. 11, pp. 55706–55715, 2023, doi: 10.1109/ACCESS.2023.3282444.

[5] T. Suzuki, D. Sato, Y. Sugaya, T. Miyazaki and S. Omachi, "Important region estimation using image captioning," in IEEE Access, vol. 10, pp. 105546–105555, 2022, doi: 10.1109/ACCESS.2022.3211260.

[6] Z. Gu and J. Jin, "Attention-guided hierarchical parsing for fine-grained person-centric image captioning," in IEEE Access, vol. 12, pp. 86293–86301, 2024, doi: 10.1109/ACCESS.2024.3416207.

[7] N. Yu, X. Hu, B. Song, J. Yang and J. Zhang, "Topic-oriented image captioning based on order-embedding," IEEE Transactions on Image Processing, vol. 28, no. 6, pp. 2743–2754, 2018, doi: 10.1109/TIP.2018.2889922.

[8] M. A. Al-Malla, A. Jafar and N. Ghneim, "Image captioning model using attention and object features to mimic human image understanding," Journal of Big Data, vol. 9, p. 20, 2022, doi: 10.1186/s40537-022-00571-w.

[9] L. Abisha Anto Ignatious, S. Jeevitha, M. Madhurambigai and M. Hemalatha, "A semantic driven CNN–LSTM architecture for personalised image caption generation," in 2019 11th International Conference on Advanced Computing (ICoAC), Chennai, India, 2019, pp. 356–362, doi: 10.1109/ICoAC48765.2019.246867.

[10] Y. Jing, X. Zhiwei and G. Guanglai, "Context-driven image caption with global semantic relations of the named entities," in IEEE Access, vol. 8, pp. 143584–143594, 2020, doi: 10.1109/ACCESS.2020.3013321.

[11] N. M. Khassaf and N. H. M. Ali, "Improving pre-trained CNN-LSTM models for image captioning with hyper-parameter optimization," Engineering, Technology & Applied Science Research, vol. 14, no. 5, pp. 17337–17343, Oct. 2024.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick, "Microsoft COCO: Common objects in context," in Computer Vision – ECCV 2014, D. Fleet, T. Pajdla, B. Schiele and T. Tuytelaars, Eds. Cham: Springer, 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.

[13] P. Young, A. Lai, M. Hodosh and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," Transactions of the Association for Computational Linguistics, vol. 2, pp. 67–78, 2014, doi: 10.1162/tacl_a_00166.

[14] M. A. Arasi, H. M. Alshahrani, N. Alruwais, A. Motwakel, N. A. Ahmed and A. Mohamed, "Automated image captioning using sparrow search algorithm with improved deep learning model," in IEEE Access, vol. 11, pp. 104633–104642, 2023, doi: 10.1109/ACCESS.2023.3317276.

[15] E. K. Wang, X. Zhang, F. Wang, T.-Y. Wu and C.-M. Chen, "Multilayer dense attention model for image caption," in IEEE Access, vol. 7, pp. 66358–66368, 2019, doi: 10.1109/ACCESS.2019.2917771.

[16] C. Wang, H. Yang, C. Bartz and C. Meinel, "Image captioning with deep bidirectional LSTMs," in Proc. 24th ACM International Conference on Multimedia, New York, NY, USA, 2016, pp. 988–997, doi: 10.1145/2964284.2964299.

[17] V. Chahkandi, M. J. Fadaeieslam and F. Yaghmaee, "Improvement of image description using bidirectional LSTM," International Journal of Multimedia Information Retrieval, vol. 7, pp. 147–155, 2018, doi: 10.1007/s13735-018-0158-y.

[18] I. Phueaksri, M. A. Kastner, Y. Kawanishi, T. Komamizu and I. Ide, "An approach to generate a caption for an image collection using scene graph generation," in IEEE Access, vol. 11, pp. 128245–128260, 2023, doi: 10.1109/ACCESS.2023.3332098.

[19] R. Sasibhooshan, S. Kumaraswamy and S. Sasidharan, "Image caption generation using visual attention prediction and contextual spatial relation extraction," Journal of Big Data, vol. 10, p. 18, 2023, doi: 10.1186/s40537-023-00693-9.

[20] S. Wang, L. Lan, X. Zhang, G. Dong and Z. Luo, "Cascade semantic fusion for image captioning," in IEEE Access, vol. 7, pp. 66680–66688, 2019, doi: 10.1109/ACCESS.2019.2917979.
[21] J. W. Bae, S.-H. Lee, W.-Y. Kim, J.-H. Seong and D.-H. Seo, "Image captioning model using part-of-speech guidance module for description with diverse vocabulary," in IEEE Access, vol. 10, pp. 45219–45229, 2022, doi: 10.1109/ACCESS.2022.3169781.
[22] D. Naik and C. D. Jaidhar, "A novel multi-layer attention framework for visual description prediction using bidirectional LSTM," Journal of Big Data, vol. 9, p. 104, 2022, doi: 10.1186/s40537-022-00664-6.
[23] W. Jiang, X. Li, H. Hu, Q. Lu and B. Liu, "Multi-gate attention network for image captioning," in IEEE Access, vol. 9, pp. 69700–69709, 2021, doi: 10.1109/ACCESS.2021.3067607.
[24] C. Wu, S. Yuan, H. Cao, Y. Wei and L. Wang, "Hierarchical attention-based fusion for image caption with multi-grained rewards," in IEEE Access, vol. 8, pp. 57943–57951, 2020, doi: 10.1109/ACCESS.2020.2981513.
[25] J. Wang and J. Feng, "Hybrid attention distribution and factorized embedding matrix in image captioning," in IEEE Access, vol. 8, pp. 154453–154460, 2020, doi: 10.1109/ACCESS.2020.3018546.
[26] L. Cheng, W. Wei, X. Mao, Y. Liu and C. Miao, "Stack-VS: Stacked visual-semantic attention for image caption generation," in IEEE Access, vol. 8, pp. 154953–154965, 2020.
[27] J. Wu, T. Chen, H. Wu, Z. Yang, G. Luo and L. Lin, "Fine-grained image captioning with global-local discriminative objective," IEEE Transactions on Multimedia, vol. 23, pp. 2413–2427, 2021, doi: 10.1109/TMM.2020.3011317.
[28] T. Jumphoo, K. Phapatanaburi, W. Pathonsuwan, P. Anchuen, M. Uthansakul and P. Uthansakul, "Exploiting data-efficient image transformer-based transfer learning for valvular heart diseases detection," in IEEE Access, vol. 12, pp. 15845–15855, 2024, doi: 10.1109/ACCESS.2024.3357946.
[29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever, "Learning transferable visual models from natural language supervision," in Proc. 38th International Conference on Machine Learning, 2021.
[30] J.-B. Alayrac, J. Donahue, P. Luc et al., "Flamingo: A visual language model for few-shot learning," in Proc. 36th International Conference on Neural Information Processing Systems (NeurIPS '22), Red Hook, NY, USA, 2022, Article 1723, pp. 23716–23736.
[31] H. Liu, C. Li, Q. Wu and Y. J. Lee, "Visual instruction tuning," in Proc. 37th International Conference on Neural Information Processing Systems (NeurIPS '23), Red Hook, NY, USA, 2023, Article 1516, pp. 34892–34916.
[32] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," arXiv preprint, arXiv:1406.1078, 2014.
[33] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink and J. Schmidhuber, "LSTM: A search space odyssey," arXiv preprint, arXiv:1503.04069, 2015.
[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NIPS), vol. 30, 2017.
[35] Z. Li, J. Wei, F. Huang and H. Ma, "Modeling graph-structured contexts for image captioning," Image and Vision Computing, vol. 129, p. 104591, 2023, doi: 10.1016/j.imavis.2022.104591.
[36] C. Bhosale, "Automatic annotation of query results from deep web database," International Journal of Engineering Sciences & Research Technology, vol. 4, no. 8, 2015.
[37] U. Karanje, B. C. S., G. Kinge, A. Magaonkar, A. Gade and J. Kankhar, "Traffic sign recognition using convolutional neural network," in Applied Soft Computing Techniques: Theoretical Principles and Practical Applications, 2025.
[38] C. S. Bhosale, P. Salve and V. Shirsath, "A review of image captioning techniques: types, deep learning advancements, and limitations," Cureus Journal of Computer Science, vol. 2, no. es44389-024-01573-w, Jul. 2025, doi: 10.7759/s44389-024-01573-w.