

A Scaled Encoder Decoder Network for Image Captioning in Hindi

Santosh Kumar Mishra¹, Sriparna Saha¹, and Pushpak Bhattacharyya²

¹ Department of Computer Science & Engineering, Indian Institute of Technology Patna, India

² Department of Computer Science & Engineering, Indian Institute of Technology Bombay, India
{santosh_1821cs03, sriparna}@iitp.ac.in¹, pb@cse.iitb.ac.in²

Abstract

Image captioning is a prominent research area in computer vision and natural language processing, which automatically generates natural language descriptions for images. Most of the existing works have focused on developing models for image captioning in the English language. The current paper introduces a novel deep learning architecture based on encoder-decoder with an attention mechanism for image captioning in the Hindi language. For encoder, decoder, and attention, several deep learning-based architectures have been explored. Hindi is the third-most spoken language globally; it is extensively spoken in India and South Asia; it is one of India's official languages. The proposed encoder-decoder architecture employs scaling in convolution neural networks to achieve better accuracy than existing image captioning methods in Hindi. The proposed method's performance is compared with state-of-the-art methods in terms of BLEU scores and manual evaluation. The results show that the proposed method is more effective than existing methods.

1 Introduction

Caption generation from images is a complex job as it necessitates object recognition and articulating the object's relationship in natural language. Caption generation is challenging in comparison to object recognition and image classification, which have been the primary research focus in computer vision. Nowadays, Deep learning-based architecture has emerged as a result of recent developments in machine translation. Recent advances in language modeling, object recognition, and image classification opened up new possibilities. A generated image caption can assist visually challenged individuals to perceive the web content (MacLeod et al., 2017). The architecture based on encoder-decoder has been widely employed to solve the

image captioning problem (Karpathy and Fei-Fei, 2015) (Anderson et al., 2018) (Feng et al., 2019). In the literature, two different approaches have been used for caption generation; the top-down approach (Bahdanau et al., 2014) (Wu et al., 2016) (Vinyals et al., 2015) (Zhou et al., 2020), (Cornia et al., 2020), and the bottom-up approach (Elliott and Keller, 2013) (Kulkarni et al., 2011) (Farhadi et al., 2010).

In this paper, We built a model of caption generation from images in the Hindi language, which is spoken throughout India, South Asia, and other parts of the world as well. It is one of the world's ancient languages and the third most spoken language globally. It originated from the Sanskrit language (Gary and Rubino, 2001). In the literature, there are just a few works on Hindi image captioning (Dhir et al., 2019; Mishra et al., 2021a,b; Singh et al., 2021). The first work was carried out in (Dhir et al., 2019), RESNET 101 (He et al., 2016), and GRU (Cho et al., 2014) is employed in the architecture. In (Mishra et al., 2021a), authors had employed various attention models. In this paper, the authors had explored several architectures with various attention. Authors of (Mishra et al., 2021b) proposed a architecture using transformer. The transformer is employed here as a decoder. This work also utilizes deep-learning based architectures for generating captions of images in the Hindi language. The key contributions of this work are as follows:

- This work is the first of its kind for image captioning in Hindi, which utilizes EfficientNet (Tan and Le, 2019) as an encoder and GRU (gated recurrent unit) (Cho et al., 2014) as a decoder with Bahdanu attention (Bahdanau et al., 2014).
- Ablation study has been conducted with various encoder-decoder and attention technique

like X-linear attention (Pan et al., 2020), Bahdanu attention (Bahdanau et al., 2014), Luong attention (Luong et al., 2015), spatial attention (Lu et al., 2017), Visual Attention (Xu et al., 2015).

- We explored various language model with the proposed architecture like Transformer (Vaswani et al., 2017), LSTM (Hochreiter and Schmidhuber, 1997) and GRU (Cho et al., 2014).
- We experimented with the newly introduced dataset for image captioning in Hindi (Mishra et al., 2021a) and showed a comparative study between model trained on Hindi dataset and post-processing model. Here, the post-processing model is trained on an English corpus that generates an English caption that is ultimately translated into Hindi. We demonstrated the efficacy of the proposed method by comparing it with the post-processing method.

2 Related Work

In the past, two approaches for image captioning have been used; the first is the top-down approach (Wu et al., 2016)(Sutskever et al., 2014) (Bahdanau et al., 2014), and second approach is an older approach i.e. bottom up approach (Elliott and Keller, 2013)(Kulkarni et al., 2011)(Farhadi et al., 2010). In the first approach, the input image is turned into words, but in the second approach, words define the many features of an image; words are joined to form an image caption. The architecture's parameters are learned in top-down methodology, which comprises end-to-end learning for caption generation.

The language model combines the object characteristics, vocabulary, visual description, and sentences, etc., in the bottom-up approach. Association of the appropriate sentence to an input image is explored in (Farhadi et al., 2010); this sentence is considered as the input image's caption. In (Elliott and Keller, 2013), a template-based method was utilized, it uses visual dependency modeling to record the links among objects.

For image captioning nowadays, the top-down method is very popular. Authors of (Mao et al., 2014) had developed the captioning architecture using the multimodal RNN to generate the caption. A probability distribution model is employed to

generate the word based on prior words and an image. The probability distribution is used to generate the image caption. It is analogous to the approach of machine translation employing encoder-decoder architecture. In (Vinyals et al., 2015), authors have used a generative model using an RNN trained to optimize the likelihood of the target sentence given an input image. Authors of (Karpathy and Fei-Fei, 2015) have proposed an image captioning model by using a combination of CNN and RNN over image region utilizing the alignment model. They used bidirectional RNN for language modeling and a structured, objective function aligning two modalities through a multimodal embedding. A language pre-training model unified version is investigated in (Zhou et al., 2020). A meshed memory transformer (Cornia et al., 2020) is utilized for image's feature extraction and language modeling; it learns a multi-level relationship between previous information and regions of the image. Authors of (Liu et al., 2020) have proposed an image captioning model using generative adversarial networks using retrieval and ensemble based approaches. The method given by (Deshpande et al., 2019) has an image captioning structure using variational generative adversarial network and variational autoencoder; the approach generates an image caption based on an image summary. The authors also used part-of-speech as a description that assists in generating the description of the image.

Most of the relevant works for image captioning in the literature are published particularly for the English. Only a limited number of attempts have been made for image caption generation in the Hindi language. In (Dhir et al., 2019), the first attempt for image captioning in the Hindi language is made. A transformer-based image captioning model has been proposed in (Mishra et al., 2021b). In (Mishra et al., 2021a), authors have investigated a variety of architectures with various attention methods for caption generation from images.

3 Proposed Methodology

We employed the encoder-decoder framework with attention for image captioning in the proposed framework (as shown in Fig 1).

3.1 Encoder-Decoder Framework

We explore the encoder-decoder based architecture for caption generation of an image. Given an image, it maximizes the correct description's probability

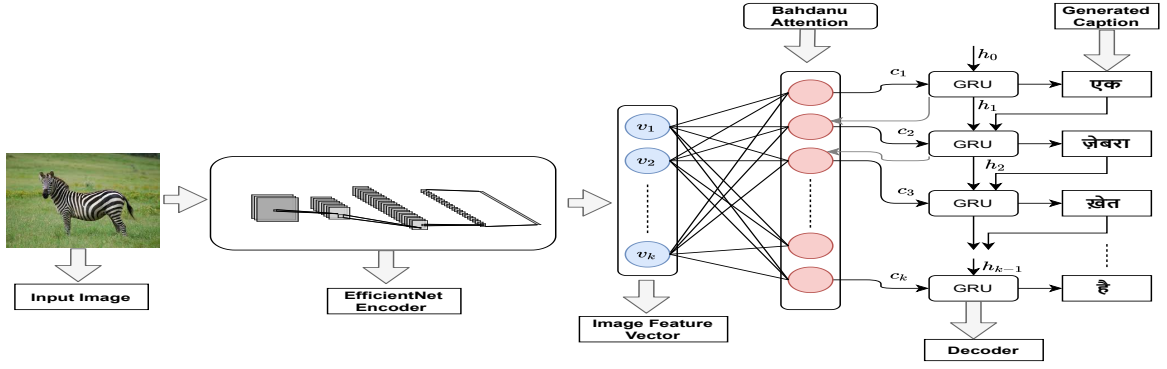


Figure 1: Network architecture of the proposed method

as follows:

$$\theta^* = \arg \max_{\theta} \sum_{(I,S)} \log p(S|I; \theta) \quad (1)$$

In the above equation, θ specifies model parameters, I represents the image and $y = y_1, y_2, \dots, y_t$ is the related caption. The generated caption y is obtained by chain law. The joint probability distribution's log-likelihood may be calculated as follows:

$$\log p(y) = \sum_{t=0}^N \log p(y_t | y_0, y_1, \dots, y_{t-1}, I) \quad (2)$$

For clarity, the model's parameter dependence has been discarded. The architecture based on RNN is as follows:

$$\log p(y_t | y_0, y_1, \dots, y_{t-1}, I) = f(h_t, c_t) \quad (3)$$

Here f is a non-linear function finds the next word's output probability. h_t and c_t represents the hidden state and context vector extracted vector of the image at t^{th} time step of recurrent neural network. The context vector c_t is an essential characteristic in this case since it offers verification throughout the caption generating process. (You et al., 2016)(Xu et al., 2015)(Mao et al., 2014)(Vinyals et al., 2015). c_t is dependent on both the encoder and decoder architectures. Its been demonstrated in prior publications; attention helps in increasing the efficiency of the image captioning model (Xu et al., 2015).

3.2 Convolutional Neural Networks as an Encoder for Feature Extraction

The proposed method encodes an input image I into a vector representation of fixed size; the en-

coded image feature sets the decoder RNN's starting state. We conducted an ablation investigation using encoders such as EfficientNet, Inception V4, and RESNET 101.

3.2.1 EfficientNet

EfficientNet is a group of convolutional neural networks(CNNs) architectures proposed by authors (Tan and Le, 2019) to optimize the accuracy for image classification given a computational cost. It employs the model scaling to find the best combination of resolution, width, and depth in CNNs. There are eight models from B0 to B7 in the EfficientNet, with each subsequent model number relating to variants with more parameters and higher accuracies. We have used the B5 model trained on ImageNet for feature extraction from the input images. More details can be found in the paper (Tan and Le, 2019).

3.2.2 RESNET 101

RESNET101 (Residual Neural Network) (He et al., 2016) is employed for image encoding and extracting features. It consists of 101 layers that are trained on the ImageNet dataset.

3.2.3 Inception V4

This CNN architecture proposed in (Szegedy et al., 2017). It has a greater number of inception components comparing Inception-V3. This is a true Inception variation with no residual links. On ImageNet's test set classification challenge, this model earned a top-5 error of 3.08%.

3.3 Attention Mechanisms

The encoder-decoder model uses a fixed-length context vector which is incapable of remembering long input sequences. The attention mechanisms resolve this problem. Attention mechanisms focus on the crucial part of the image while generating the

caption. Proposed encoder-decoder model makes use of, recently introduced X-Linear attention (Pan et al., 2020), Spatial Attention (Lu et al., 2017), Visual Attention (Xu et al., 2015), Luong Attention (Luong et al., 2015), and Bahdanau Attention (Bahdanau et al., 2014).

3.3.1 Bahdanau Attention

This architecture introduced in (Bahdanau et al., 2014) is a well known architecture for sequence to sequence model. This is a kind of additive attention; here context vector is calculated as follows:

$$c_t = \sum_{i=1}^N \alpha_{ti} v_i \quad (4)$$

the weight α_{ti} for each feature v_i is determined as :

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{i=1}^N \exp(e_{ti})} \quad (5)$$

where

$$e_{ti} = f(h_t, v_i) \quad (6)$$

Here, the proposed feed forward neural network (Bahdanau et al., 2014) is denoted by f which is jointly trained on all parameters, v_i denotes image feature, h_t is RNN's hidden state at time step t and N is the generated caption's length.

3.3.2 Luong Attention

This attention mechanism (Luong et al., 2015) is commonly referred as multiplicative attention, which is built upon the Bahdanu attention. Here, c_t denotes the model vector is determined as follows:

$$c_t = \sum_{i=1}^N \alpha_{ti} v_i \quad (7)$$

In this case, the weight α_{ik} is determined for each feature v_k as :

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{i=1}^N \exp(e_{ti})} \quad (8)$$

where

$$e_{ti} = h_t \times w \times v_i \quad (9)$$

e_{ti} represents content based function (Luong et al., 2015), v_i is the image feature vector, N is the length of the generated caption, h_t denotes hidden states at time step t , and w represents the learnable parameters.

3.3.3 Visual Attention

Authors of (Xu et al., 2015) demonstrated an attention technique for focusing on the appropriate portion of the image while generating a caption. Here, the context vector is computed using:

$$e_{ti} = f(v_i, h_t) \quad (10)$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{i=1}^N \exp(e_{ti})} \quad (11)$$

$$c_t = \phi(v_i, \alpha_{ti}) \quad (12)$$

Here, d dimensional feature vectors of different parts of the images are $V = [v_1, v_2, \dots, v_k]$, $v_i \in R^d$ is the spatial image feature. h_t denotes the hidden state of recurrent neural network at t^{th} time step. α_{ti} denotes the weight which is computed for each image feature vector, v_i , at every time step by a proposed attention architecture, f (Xu et al., 2015). It employs a multilevel perceptron applied on hidden state, h_t , and context vector, c_t . The function ϕ returns a single vector corresponding to their weights, further h_t and c_t are jointly utilized to anticipate the succeeding word as given in Equation 3.

3.3.4 Spatial Attention

This is generated from the residual network (He et al., 2016). This mechanism utilizes residual connection (He et al., 2016); authors have introduced a new technique of determining the context vector, it is regarded as the present hidden state's residual visual information.

$$c_t = g(V, h_t) \quad (13)$$

Here attention function is represented by g and $V = [v_1, v_2, \dots, v_k]$, $v_i \in R^d$ represents d dimensional feature vector of image. v_i and h_t represent parts of image and RNN's hidden state at time step t , respectively.

3.3.5 X-Linear Attention

The conventional attention module primarily uses first-order interaction for image captioning, which has limited multi-modal reasoning capacity. Second-order interaction (bilinear pooling) has been demonstrated to be helpful in visual recognition by the authors of (Gao et al., 2016), and (Yu et al., 2018). This mechanism has been utilized by the authors of (Kim et al., 2018), and (Fukui et al., 2016) for visual question answering. X-linear attention uses bilinear pooling that boosts the attended

feature’s capacity of representation by utilizing the higher-order interaction between uni-modal and multi-modal features.

Let’s suppose that $Q \in R^{D_q}$ represent the query, $K = \{k_i\}_{i=1}^N$ represent the keys and $V = \{v_i\}_{i=1}^N$ denote the set of values, where $v_i \in R^{D_v}$ and $k \in R^{D_k}$ are i^{th} value and key pair, respectively. Lower rank bi-linear pooling is used by X-linear attention to obtain a query-key representation, $B_i^k \in R^{D_B}$, between query, Q , and each key, k_i .

$$B_i^k = \sigma(W_k k_i \odot \sigma(W_q^k Q)) \quad (14)$$

$W_q^k Q$ and $W_k \in R^{D_B \times D_k}$ are the embedding matrices, sigma (σ) depicts the relu activation function and \odot is the multiplication of elements. Here B_i^k specifies learned bi-linear query representation, and it represents an interaction between the key and query on a second-order level.

Furthermore, two types of bi-linear distributions of attention are calculated to aggregate both channel-wise and spatial information across all values. Two embedding layers are used to get the distribution of spatial attention. The bi-linear query representation is then projected into corresponding attention weight using a softmax layer.

$$B_i'^k = \sigma(W_B^k B_i^k) \quad (15)$$

$$b_i^s = W_b B_i'^k \quad (16)$$

$$\beta^s = softmax(b^s) \quad (17)$$

Where W_b and $W_B^k \in R^{D_c \times D_B}$ are the embedding matrices. $B_i'^k$ represents bi-linear query-key representation and b_i^s denotes the i^{th} element in b^s . Each of the elements β_i^s in β^s represents a key/value pair’s attention weight. Further squeeze-excitation (Hu et al., 2018) is performed over all transformed bi-linear query representations, $\{B_i'^k\}_{i=1}^N$, for attention measurement in channel wise manner. The squeeze operation uses average pooling to aggregate all of the modified bi-linear key and query representations, yielding a global channel descriptor, \bar{B} , as follows:

$$\bar{B} = \frac{1}{N} \sum_{i=1}^N B_i'^k \quad (18)$$

Further, channel wise attention distribution is derived by excitation operation, β^c , by using the

self gating with sigmoid activation function over the global channel descriptors, \bar{B} .

$$b^c = W_e \bar{B} \quad (19)$$

$$\beta^c = sigmoid(b^c) \quad (20)$$

The embedding matrix is $W_e \in R^{D_B \times D_c}$ in this case. Finally, the X-Linear attention module produces the attended features of images by combining improved bi-linear values with channel-wise and spatial attention to form the attended feature.

$$\hat{v} = F_{X-linear}(K, V, Q) = \beta^c \odot \sum_{i=1}^N \beta_i^s \beta_i^v \quad (21)$$

$$B_i^v = \sigma(W_v v_i) \odot \sigma(W_q^v Q) \quad (22)$$

Where $W_v \in R^{D_B \times D_v}$ and $W_q^v \in R^{D_B \times D_q}$ are the embedding matrices, B_i^v represents the bi-linear pooling’s enhanced values on query, Q , and value, v_i . In contrast to the traditional attention framework that utilizes only the first-order interaction, the X-linear attention model utilizes the second-order interaction via bi-linear pooling. Therefore, it has more representative attended features than the traditional attention framework.

3.4 Decoder for Language Modeling

We have used various decoder models for ablation study and to determine the best possible architecture. The language modeling RNN has the challenge of exploding and vanishing gradient (Hochreiter and Schmidhuber, 1997). This problem can be resolved employing gated recurrent unit (Cho et al., 2014), and Long Short-Term Memory (Hochreiter and Schmidhuber, 1997). We have included a bi-directional variation in addition to the uni-directional GRU and LSTM, which enables the networks to have forward and backward sequence information. We have also incorporated the transformer (Vaswani et al., 2017) as a decoder; apart from attention, to enable optimization easier and quicker, it employs positional encoding, residual connection, and layer normalization.

4 Experimental Setup

This section covers the methods employed to create the dataset and evaluate the proposed methodology.

4.1 Dataset

The authors of (Mishra et al., 2021a) generated the Hindi variant of the MSCOCO dataset. This is a popular dataset for caption generation from images. In this dataset, each image has five captions. There are 82573, 811, 811 images for training, testing, and validation. The captions in the training set, validation set, and test set are about 4 lakh, 4000, and 4000, respectively. Despite the fact Google Translate is employed for the translation, the following difficulties have been experienced when translating from English to Hindi:

- Because Google Translate lacks a system to assess the context of the statement, the context of the translated caption is lost during translation.
- In certain cases, Google Translator’s translation is grammatically imprecise.
- Google Translator’s accuracy is not standardized because it depends on the source and target languages.

Therefore, human annotators are employed to correct Google translated sentences to remove errors. The inter-annotator agreement was 87% between two annotators. Figures 2 and 3 display a sample from the dataset that was created.



Figure 2: Example Image for Dataset Preparation

English Caption	Google Translated Caption in Hindi	Corrected Caption in Hindi
A long empty, minimal modern skylit home kitchen.	एक लंबा खाली, न्यूनतम आधुनिक स्काईलाइट होम किचन।	एक लंबा खाली, छोटा आधुनिक रोशन दान युक्त घर की रसोई।
A picture of a modern looking kitchen area.	आधुनिक दिखने वाले रसोई क्षेत्र की एक तस्वीर।	आधुनिक दिखने वाले रसोई क्षेत्र की एक तस्वीर।
A narrow kitchen ending with a chrome refrigerator.	क्रीम रेफ्रिजरेटर के साथ समाप्त होने वाली एक संकीर्ण रसोई।	क्रीम रेफ्रिजरेटर के साथ समाप्त होने वाली एक संकीर्ण रसोई।
A narrow kitchen is decorated in shades of white, gray, and black.	एक संकीर्ण रसोईघर सफेद, ग्रे और काले रंग के रंगों में सजाया गया है।	एक संकीर्ण रसोईघर सफेद, भूरा और काले रंग के रंगों में सजाया गया है।
a room that has a stove and a icebox in it.	एक कमरा जिसमें एक स्टोव और एक आइसबॉक्स है।	एक कमरा जिसमें एक चूल्हा और एक बर्फ रखने का डिब्बा है।

Figure 3: Example of Dataset Preparation

4.2 Evaluation Metric

We employed the BLEU score (Papineni et al., 2002), a standard evaluation measure used in image captioning and machine translation etc.

4.3 Hyperparameters Used

EfficientNet extracts feature from $224 * 224$ input images and transform them into $49 * 512$ feature vectors. Embedding layer size is 512 neurons, 0.4 dropouts are employed to prevent over-fitting. The batch size is fixed to 128, and the epochs are set to 15. Softmax cross-entropy is employed as a loss function. The Adam optimizer with a $4e - 4$ learning rate is used for optimization. It takes 14 hours to train; a caption for the image needs around 30 to 40 seconds to generate.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Proposed Methodology (EN + BA + GRU)	67.3	48.5	33.1	22.0
EN + BA + LSTM	66.7	47.6	32.4	21.8
EN + VA + GRU	66.5	47.3	32.1	21.4
EN + XLA + GRU	66.8	47.8	32.1	21.0
EN + XLA + LSTM	67.2	48.0	32.7	21.9
EN + XLA + Bi-GRU	66.1	47.0	31.5	20.6
EN + XLA + Bi-LSTM	66.7	47.6	32.1	21.2
IV4 + SA + GRU	55.7	38.4	25.9	16.8
EN + Trans	62.7	43.7	28.8	18.2
RN101 + SA + LSTM	56.4	38.8	26.3	17.2
IV4 + SA + LSTM	56.3	38.4	25.8	16.9
EN + LA + GRU	67.0	48.4	32.4	21.0
EN + SA + GRU	66.0	46.7	31.4	20.5
Mishra et al. (Mishra et al., 2021a)	67.0	47.8	31.9	21.2
Mishra et al. (Mishra et al., 2021b)	62.9	43.3	29.1	19.0
Dhir et al. (Dhir et al., 2019)	57.0	39.1	26.4	17.3

Table 1: The score obtained with various architectures and comparison with existing methods. Here Trans, Bi-LSTM, LSTM, Bi-GRU, GRU, SA, LA, BA, VA, XLA, IV4, RN101, and EN represents Transformer, Bi-directional Long Short-Term Memory, Long Short-Term Memory, Bi-directional Gated Recurrent Unit, Gated Recurrent Unit, Spatial Attention, Luong Attention, Bahdanu Attention, Visual Attention, X-Linear Attention, Inception V4, RESNET101, and EfficientNet



C1: एक सफेद प्लेट पर एक हॉट डॉग
C2: A hot dog on a table
C3: एक मेज पर एक गर्म कुत्ता

C1: Generated caption based on model trained on Hindi dataset
C2: Generated caption based on model trained on English dataset
C3: Generated caption based on model trained on English dataset then translated into Hindi

Figure 4: Captions generated by different models of test images. Generated caption, Gloss and Transliteration are denoted by I,II, and III.

5 Results and Discussions

A comprehensive overview of obtained results and generated captions are discussed in this section.

Score	Adequacy (Meaning)	Fluency(Meaning)
0	Poor: In the caption generated, none of the information is retained.	Poor: The Hindi caption generated is incomprehensible.
1	Bad: There is little information retained in the caption generated.	Bad: The Hindi caption generated is dis-fluent.
2	Moderate: Much of the information in the caption generated are retained.	Moderate: The Hindi captions generated are like non-native Hindi captions.
3	Good: Most of the information in the caption produced is retained.	Good: In terms of Hindi grammar rules, the generated Hindi captions are good.
4	Excellent: In the produced caption, all of the information are retained.	Excellent: Hindi captions generated are correct in terms of Hindi grammar rules.

Table 2: Adequacy and fluency measurement scale







		
(a): I- एक बिल्ली एक सोफे पर लेटी हुई है II- One cat one couch on lying down III- Ek bille ek sofe par leti hui hai	(b): I- एक आदमी एक सेल फोन पर बात कर रहा है II- One man one cell phone on talking III- Ek adami ek cell phone par bat kar raha hai	(c): I- एक मेज पर फलों और सब्जियों का एक गुच्छा II- One table on fruits and vegetables of one bunch III- Ek mej par phalo aur sabjiyo ka ek guchha
		
(d): I- एक आदमी एक टेनिस कोर्ट पर टेनिस खेल रहा है II- One man one tennis court on tennis playing III- Ek adami ek tennis court par tennis khel raha hai	(e): I- हाथियों का एक समूह एक खेत में घूम रहा है II- Elephant of one group one field in moving III- Hathiyo ka ek samooch ek khet me ghoom raha hai	(f): I- एक पीले रंग की बस जो सड़क पर खड़ी है II- One yellow color of bus which road on parked III- Ek peele rang ke bus jo sadak par khadee hai

Figure 5: Generated qualitative results on test images. Generated caption, Gloss and Transliteration are denoted by I,II, and III.


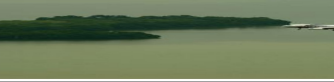
		
(a): I- एक जेबरा एक खेत में खड़ा है II- One zebra one field in standing III- Ek zebra ek khet me khada hai	(b): I- घास में एक मैदान में खड़े मवेशियों का झुंड II- Grass in one ground in standing of herd of cattle III- Ghas me ek maidan me khade maveshiyon ka ek jhund	(c): I- एक झील के ऊपर एक बड़ा सफेद नाव पानी में है II- One lake of above one big white boat in water III- Ek jheel ke upar ek bada safed naav pani me hai

Figure 6: Qualitative results to show error analysis on test images. Generated caption, Gloss and Transliteration are denoted by I,II, and III.

5.1 Comparisons with existing methods for Image Captioning in Hindi

The following works have been undertaken for image captioning in Hindi as per our understanding:

- In (Dhir et al., 2019), author have proposed the architecture for caption generation, where they had used RESNET 101 (He et al., 2016) and GRU (Cho et al., 2014).
- A transformer-based architecture introduced in (Mishra et al., 2021b), where transformer is utilized for language modeling.
- (Mishra et al., 2021a) investigates a variety of architectures with various attention methods for Hindi image captioning.

As a result, we evaluated our technique to these approaches, Table 1 show that our approach beats the existing method and baselines of the ablation study.

5.2 Qualitative Analysis

We cover the qualitative examination of our approach using test images in this section. The captions for the test images that were generated are shown in Fig 5. Gloss annotations and transliterations are added for non-Hindi speakers; they help comprehend the captions in Hindi. It is obvious that the produced captions are mostly accurate and can appropriately signify the items and activities depicted in the images.

5.3 Quantitative Analysis

The efficiency of the proposed method was assessed using BLEU scores, as can be seen in Table 1. This table depicts that our method surpasses existing approaches considering BLEU. This demonstrates the effectiveness of our approach.

5.3.1 Human Evaluation Based on Adequacy and Fluency

These metrics are widely employed in various natural language processing problems, for-instance summarization, question-answering, and machine translation. Adequacy measures information retained in the caption generated and fluency tests generated caption in terms of grammatical norms. These metrics were evaluated on a scale of 0 to 4 (as indicated in Table 2). Two human annotators have accomplished this task with an agreement of 87% between them.

The generated captions of two approaches have been measured here:

- The approach employs a dataset including Hindi corpora in the training phase. The trained model generates captions in Hindi. This yields a score of 3.112 for adequacy and 3.233 for fluency.
- Another approach uses an English corpus for training and generates captions in English. The Google Translator is being used to convert the produced English caption into Hindi. This yields a score of 2.142 for adequacy and 2.761 for fluency.

Our methodology is superior to the post-processing procedure (The Hindi captions are formed by translating the English captions generated by the trained model with the English corpus.). The generated captions are presented in Fig 4, and the result is that the model trained on the Hindi dataset beats the post-processing procedure, which highlights the need for a Hindi dataset.

5.4 Error Analysis

There are some challenges for the image captioning framework that results in errors during caption generation (as shown in Fig 6). These challenges could be categorized as following:

- **Recognition of activity:** As can be observed in Fig 6 (a), a zebra is really sprinting, yet the model predicted that it would be standing.

This might be because the bulk of the images in dataset has a standing zebra.

- **Objects counting:** There are two animals in the picture in 6 (b), however, the model predicted 'herd of cattle.' This might be due to trained CNN's inability to detect the number of objects.
- **Occlusion:** It occurs when objects are partially visible or so near that the machine learning model can't recognize them. As can be observed in Fig 6 (c), In the caption, the model predicted 'boat' rather than 'aeroplane.'

6 Conclusion and Future Work

We present a novel approach for caption generation from images in Hindi that employs an encoder-decoder model based on EfficientNet and GRU, as well as attention techniques. We use EfficientNet as an encoder because its efficacy outperforms state-of-the-art CNNs for image classification and feature extraction. We use a gated recurrent unit as a decoder for language modeling as it is less computationally expensive and it achieves state-of-the-art efficacy for language modeling. Further, the use of Bahdanau attention makes the system robust. Aside from that, we undertake an ablation analysis to find the ideal architecture. The proposed methodology could be expanded for image-to-paragraph generation and dense image captioning.

Acknowledgments

Dr. Sriparna Saha gratefully acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya Ph.D. Scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia) for carrying out this research.

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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

- Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G. Schwing, and David Forsyth. 2019. Fast, diverse and accurate image captioning guided by part-of-speech. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Rijul Dhir, Santosh Kumar Mishra, Sriparna Saha, and Pushpak Bhattacharyya. 2019. A deep attention based framework for image caption generation in hindi language. *Computación y Sistemas*, 23(3).
- Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302.
- 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 *European conference on computer vision*, pages 15–29. Springer.
- Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Unsupervised image captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.
- Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. 2016. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326.
- Jane Gary and Carl Rubino. 2001. Facts about the world’s languages: An encyclopedia of the world’s major languages.
- 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.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks, 7132–7141. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT.
- Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.
- Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574.
- Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*. Citeseer.
- Junhao Liu, Kai Wang, Chunpu Xu, Zhou Zhao, Ruifeng Xu, Ying Shen, and Min Yang. 2020. Interactive dual generative adversarial networks for image captioning. In *AAAI*, pages 11588–11595.
- Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383.
- Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.
- Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people’s experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5988–5999. ACM.
- Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*.
- 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.
- 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.

- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- Alok Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2021. An encoder-decoder based framework for hindi image caption generation. *Multimedia Tools and Applications*, pages 1–20.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information systems*, pages 3104–3112.
- Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Mingxing Tan and Quoc V Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- 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.
- Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 203–212.
- 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.
- Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659.
- Chaojian Yu, Xinyi Zhao, Qi Zheng, Peng Zhang, and Xinge You. 2018. Hierarchical bilinear pooling for fine-grained visual recognition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 574–589.
- Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, pages 13041–13049.