

A Deep Learning based Framework for Image Paragraph Generation in Hindi

Santosh Kumar Mishra¹, Sushant Sinha¹, Sriparna Saha¹, Pushpak Bhattacharyya²

Indian Institute of Technology Patna¹, Indian Institute of Technology Bombay²

{santosh_1821cs03, sushant_1901cs62, sriparna}@iitp.ac.in¹, pb@cse.iitb.ac.in²

Abstract

Generating a paragraph from an image is a complex task that requires object and activity recognition. It is now feasible because of recent advances in image captioning. Summarizing an image into a single sentence can only provide a superficial description of the visual information included within the image. This problem can be solved by generating a detailed and coherent description of the input image. Most existing works on image-to-paragraph generation have been accomplished in English. We propose a novel way of generating a paragraph from an image in Hindi. The world's third most spoken language and one of India's official languages, Hindi, is extensively spoken throughout India and South Asia. We construct a new dataset for an image-to-paragraph generation in Hindi. We employ a hierarchical recurrent neural network (HRNN) for language modeling and an object detection model to decompose the input image by distinguishing regions of interest and objects. The performance of the proposed methodology is compared with other baselines in terms of BLEU, CIDER, and METEOR scores, and the obtained results show that the proposed method outperforms them.

1 Introduction

Human perception is dependent on vision and language. These are the most effective means of interacting with others. In our everyday lives, we are exposed to a large number of images from various sources such as advertisements, news stories, and the internet, among others. As a result, a technique to define the visual information and activity inside an image is essential. A major challenge in computer vision and natural language processing is developing a system that can explain visual objects and their relationships in natural language. A computer-generated image description may aid visually im-

paired people in comprehending online information (MacLeod et al., 2017). Recent advances in artificial intelligence, computer vision, and natural language processing have enabled image or video captioning to explain the visual content of an image or video (Vinyals et al., 2015) (Vinyals et al., 2016) (Anderson et al., 2018), (Anderson et al., 2018) Because of the availability of large datasets (Young et al., 2014) (Lin et al., 2014) that combine images with natural language descriptions, it is now possible to generate sentences to interpret the images. Though the effectiveness of these solutions is promising, they all have one big flaw: they all fail to capture the subtleties inside an image or video.

The generated caption in the image or video captioning procedure is a single sentence of around 20 words that anticipates just the essential observations within an image. Capturing all of the visual information and semantics included inside an image or video is inadequate. This problem can be solved by generating a paragraph description from an image that has comprehensive visual information. When compared to single-sentence captioning, the generation of a paragraph from images is a relatively new research area. Visual Genome, an image-to-paragraph generation dataset introduced by (Krause et al., 2017) plays an important role in the image-to-paragraph generation. This data set can be used for training any machine learning classifier to develop an image captioning model. But, machine learning models can not generate a precise paragraph for a variety of images. They generate repetitive sentences very often. To resolve this issue, most researchers have used hierarchical LSTMs, which generate separate words and sentence topics. Recent research works on image-to-paragraph generation (Johnson et al., 2016)(Krause et al., 2017)(Yu et al., 2016) (Liang et al., 2017) provide a bigger narrative while generating the description for an image or video. Image-to-paragraph generation is a challenging task that re-

quires understanding images and language modeling.

Previously, image-to-paragraph generation work was limited to the English language. In this paper, we present the first image-to-paragraph model in the Hindi language. As encoder and decoder, we use Faster R-CNN and hierarchical recurrent neural network, respectively. This work has made the following significant contributions:

- This is the first effort of its kind for paragraph generation from an image in Hindi. We use Faster R-CNN as an encoder to decompose the input image by recognizing distinct objects and regions of interest. Furthermore, the features of these regions are combined to form a rich representation of image semantics. As a decoder, a hierarchical neural network composed of sentence RNN and word RNN is employed; it uses these rich representations for language modeling.
- We present a novel Hindi dataset for paragraph generation from an image by translating a well-known Visual Genome dataset (Krause et al., 2017). Using Google Translate, we translate the whole corpus from English to Hindi. Furthermore, human annotators correct the translation according to Hindi grammatical rules, which requires a substantial amount of human effort and time.

2 Related Works

Many studies have been carried out in the past to combine visual and textual data. It has been accomplished in a variety of methods throughout the literature. Some researchers have addressed this as a ranking issue, using the image as input to identify the appropriate caption from the dataset and vice versa (Farhadi et al., 2010) (Hodosh et al., 2013).

An encoder-decoder architecture is used in nearly all recent image captioning models in the literature. The encoder is a CNN architecture pre-trained for image classification, while the decoder is mainly an LSTM or GRU as proposed by (Vinyals et al., 2015). In most cases, a convolutional neural network (CNN) is used to generate an encoding of the given source image. After that, the image encoding is put into an RNN, which selects a collection of

words (from a dictionary) that match the most with the image encoding. In (Xu et al., 2015), authors have employed RNNs as a decoder with an attention mechanism for caption generation from images; this mechanism focuses on the relevant parts of the image while generating the caption. Using a faster R-CNN (Ren et al., 2015) object detection model, in (Anderson et al., 2018), bottom-up and top-down attention mechanisms are introduced for caption generation from images. A modified encoder-decoder model with a guiding network is also utilized for image captioning (Jiang et al., 2018), the data to the decoder at each time step is the output of the guiding network. An unsupervised method of learning for image caption generation is introduced in (Feng et al., 2019), the proposed model did not employ image and sentence pairs for image captioning. A meshed memory transformer network is introduced for image captioning in (Cornia et al., 2020), it uses a multi-level representation of the region’s relationship with prior information. An image captioning model based on ensemble generation and retrieval using generative adversarial networks is explored in (Liu et al., 2020). A language pre-training model’s unified version is developed in (Zhou et al., 2020); it accomplishes language modeling based on the shared transformer network. The captions produced by the above methods are generally brief, comprising only a single phrase of no more than 20 words.

Intuitively, image to paragraph generation appears to be similar to image captioning: given an image, generate a written description of its content (Krause et al., 2017). The inventiveness in the textual description, on the other hand, is essential for the image to paragraph generation. The image-to-paragraph generation framework, in particular, is intended to generate a paragraph consisting of five or six sentences that describe the image in more detail. Furthermore, a seamless transition between the sentences of the paragraph’s phrases is required. Authors of (Johnson et al., 2016) proposed a method for producing comprehensive captions. A focus on a story theme underlying a specific image was lacking while producing engaging words separately. In [21], the authors proposed a method to deal with this issue. A two-stage hierarchy of RNNs is used in their language model. Given a visual representation of semantically significant areas in an image, the first

RNN level generates a sentence vector. This subject vector is converted into a sentence at the second RNN level. They released the first large-scale image-to-paragraph generation dataset, a subset of the Visual Genome dataset, as well as many paragraph captioning algorithms. The author of (Liang et al., 2017) added a third (paragraph-level) LSTM to this model (Krause et al., 2017), as well as adversarial training. Three LSTMs, two attention mechanisms, a phrase copy mechanism, and two adversarial discriminators are all included in their model (RTT-GAN).

Previously, the majority of studies were undertaken simply for the generation of paragraphs from images in English. To the best of our knowledge, no attempt has been made to generate paragraphs from images in Hindi. Our methodology is the first of its kind that generates paragraphs from images in Hindi.

3 Proposed Method

The proposed method takes an image as input and generates a natural language paragraph description of the image, making use of the compositional structure of both images and paragraphs (as illustrated in Fig 1). It deconstructs the input image by recognizing objects inside and other regions of interest and then combines features from all of these components to construct a pooled representation that reflects the image semantics.

A hierarchical recurrent neural network comprising two levels: a sentence RNN and a word RNN, takes this feature vector as input. The image features are sent to the sentence RNN, which then determines how many sentences to generate in the resulting paragraph and generates an input topic vector for each sentence. The word RNN generates the words of a single sentence given this topic vector. This section has a brief explanation of each of these modules, which are as follows:

3.1 Detection of Regions using Region Proposal Network

The proposed method uses a region proposal network (RPN) to detect regions of interest (ROI) as introduced in (Ren et al., 2015). It takes an input image of dimension $3 \times H \times W$ and finds regions of

interest, and generates a $D = 4096$ feature vector for each region. H and W are the height and width of the image, respectively. A convolutional neural network using the VGG-16 network processes the input image. It generates a feature map, which is subsequently processed by a region proposal network that regresses from a group of anchors. The region detector is trained in an end-to-end manner (Ren et al., 2015) for object recognition and for dense image captioning as well (Johnson et al., 2016), given a dataset consisting of captions with areas of interest. The region detector is trained for object detection (Ren et al., 2015) is utilized for the dense image captioning model (Johnson et al., 2016), using a dataset of images and corresponding ground ROI. We employ a region detector trained for dense caption generation of images on the visual genome dataset (Krishna et al., 2016), utilizing a publicly available implementation of (Krause et al., 2017) because the paragraph description does not contain annotated grounding to ROI (region of interest).

3.2 Region Pooling

The region proposal network detects different regions and generates a set of vectors $v_1, \dots, v_M \in R^D$, denoting various regions in the input images. These vectors are aggregated into a pooled vector $v_p \in R^D$, which describes the content of an image. Pooled vector v_p is computed using element-wise maximum as follows:

$$v_p = \max_{i=1}^M (W_{pool} v_i + b_{pool}) \quad (1)$$

Here $W_{pool} \in R^{D \times D}$ is a learned projection matrix, and bias $b_{pool} \in R^D$ is the bias.

3.3 Language Modeling Hierarchical Recurrent Neural Network

An HRNN based language model consists of two components: word and sentence RNN. The number of sentences to be generated is decided by the sentence RNN; it generates a topic vector of dimension P . A hierarchical neural language model is given the pooled region vector $v_p \in R^D$ as input. The word RNN produces words for a sentence given a topic vector. For both word RNN and sentence RNN, we use the conventional LSTM architecture (Hochreiter and Schmidhuber, 1997).

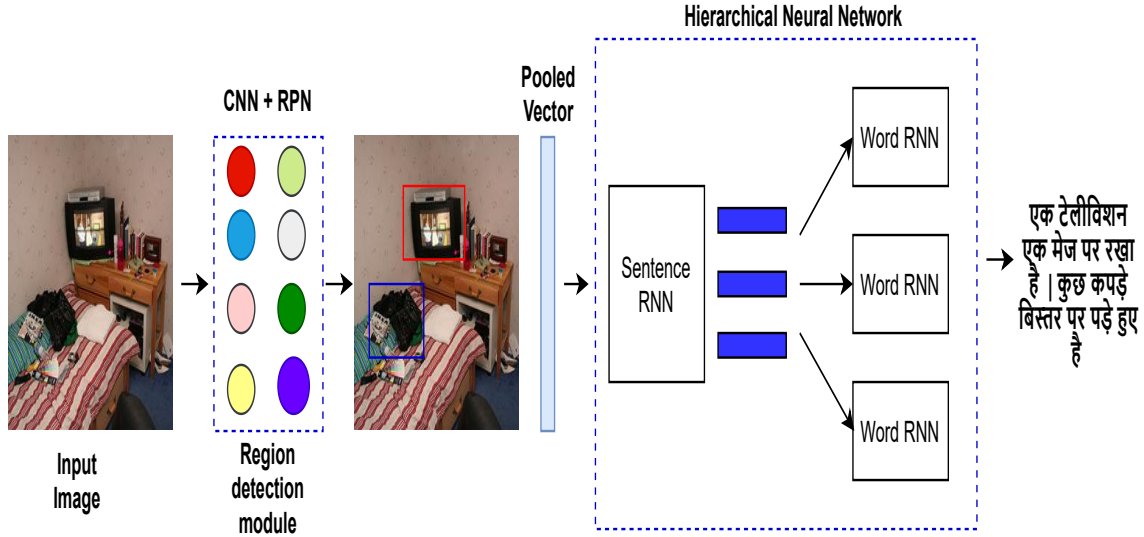


Figure 1: Architecture of the proposed method

Sentence RNN is an LSTM consisting of a single layer; its cells and hidden states are initialized with zero with a hidden size $H = 512$. At each time step, the pooling region vector v_p is given as an input to the sentence RNN, and it produces a series of hidden states, H , one for each sentence in the paragraph. The word RNN consists of two-layered LSTM with a hidden size of $H = 512$; it generates the words of the sentence given the topic vector, $t_i \in R^p$. The subject vector and a special START token are the RNN's initial and second inputs, respectively, while successive inputs are learned embedding vectors for the sentence's words. A unique END token indicates the end of a sentence, and the hidden state of the final LSTM layer is utilized to forecast a distribution across the words in the vocabulary at each timestep. Following the generation of the words for each sentence by word RNN, the sentences are combined to form the resulting paragraph.

3.4 Training Procedure

This section provides a detailed description of the training procedure. The training data is made up of pairs (x, y) where x and y represent an image and a corresponding ground truth paragraph description, respectively. Here, y consists of S number of sentences. The i^{th} sentence has N_i words and y_{ij} denotes the j^{th} word of the i^{th} sentence. Sentence

RNN is unrolled for S sentences after a pooled region vector v_p is computed for an image. For each sentence, the sentence RNN generates the probability distribution p_i over the $CONTINUE, STOP$. Here, $CONTINUE, STOP$ are the special keywords to determine when to stop or continue generating the sentences in the paragraph.

Training loss $L(X, Y)$ is the weighted summation of word loss L_{word} and sentence loss $L_{sentence}$. It is defined as follows:

$$L(x, y) = \lambda_{sent} \sum_{i=1}^S L_{sent}(p_i, I[i = S]) + \lambda_{word} \sum_{i=1}^S \sum_{j=1}^{N_i} L_{word}(p_{ij}, y_{ij}) \quad (2)$$

Here, the sentence RNN generates the sentences until it reaches S_{max} or stopping probability, $p_i(STOP)$, exceeds a threshold, T_{stop} . Here, values of the above parameters are as follows; $T_{stop} = 0.5$, $S_{Max} = 6$ and $N_{MAX} = 50$. We also incorporate self-critical sequence training in the above architecture to enhance the diversity in the paragraph generation (Rennie et al., 2017).

4 Experimental Setup

4.1 Dataset

We construct a dataset for the task of paragraph generation from images in Hindi by translating the well-known Stanford image to paragraph generation dataset (Krause et al., 2017) from English to Hindi,¹. It has a total of 19,551 images taken from the MSCOCO (Lin et al., 2014) and Visual Genome (Krishna et al., 2016) data sets. The dataset is divided into 3 parts: 14,575 training images, 2,487 validation images, and 2,489 testing images. Initially, all of the English captions were translated into Hindi by Google Translator; the following issues were encountered while translating from English to Hindi:

- Since Google Translate lacks a technique for adding sentence context, the translated caption’s meaning is lost during translation.
- In certain cases, Google Translator’s translation is grammatically incorrect.
- The Google Translator accuracy is not uniform as it is dependent on pairings of source and target languages.

As a result, human annotators are employed to correct Google-translated sentences to remove errors. The inter-annotator agreement was 87% between the two annotators. A sample example from the dataset is shown in Fig 2 and Fig 3.

Though we can get the paragraph for an image in Hindi by translating the English paragraph generated by a model trained for English, the resultant caption lacks adequacy and fluency, as shown by the authors of (Mishra et al., 2021a) (Mishra et al., 2021b). This demonstrates the need of constructing a Hindi dataset for image-to-paragraph generation.

4.2 Evaluation Metrics

We evaluate the proposed methodology using BLEU (Papineni et al., 2002), CIDEr-D (Vedantam et al., 2015), and METEOR (Denkowski and Lavie, 2014).

¹dataset will be released on acceptance

4.3 Hyper-parameters Used

The proposed architecture incorporates two layers of LSTM with a dimension of 512. The dimension of feature pooling is 1024. Stochastic gradient descent with the Adam optimizer (Kingma and Ba, 2014) is used for training. Values of λ_{sent} and λ_{word} are set to 5.0 and 1.0, respectively. The model has been trained for 30 epochs, which takes approximately 12 hours of training.

5 Results and Discussion

This section covers the detailed discussions of the results and analysis. We carried out the experimentation on the introduced image-to-paragraph generation dataset in Hindi.

5.1 Comparative baselines for image to paragraph generation

To the best of our understanding, no work has been done on paragraph generation from images in the Hindi language. Therefore, we create our own baselines, which are as follows:

- **Baseline -1:** In this baseline, top-down attention (Anderson et al., 2018) is incorporated with Faster-R CNN (Ren et al., 2015) and bi-LSTM (Hochreiter and Schmidhuber, 1997), here we explore the bi-directional LSTM for language modeling.
- **Baseline-2:** In this baseline, adaptive attention (Lu et al., 2017) is incorporated with Faster-R CNN (Ren et al., 2015) and LSTM (Hochreiter and Schmidhuber, 1997).
- **Baseline-3:** In this, we explore adaptive attention (Lu et al., 2017) with Faster-R CNN (Ren et al., 2015) and LSTM (Hochreiter and Schmidhuber, 1997) with Maxout (MO) activation function (Goodfellow et al., 2013).

5.2 Qualitative Evaluation

This section shows a qualitative evaluation of the proposed methodology on test images. The generated paragraph for the test image is shown in Fig 4. We include the transliteration and gloss annotation so that non-Hindi speakers can grasp the meanings of the captions. It can be seen from the Fig 4 that the



Figure 2: A sample image from the dataset created in Hindi

Original paragraph in English	Google translated paragraph in English	Corrected paragraphs in Hindi
A baseball game is being played. The batsman is wearing a red jersey. Two people are standing behind him. More teammates are sitting in the dugout.	बेसबॉल खेल खेला जा रहा है। बल्लेबाज ने लाल जर्सी पहनी हुई है। उसके पीछे दो लोग खड़े हैं। डगआउट में टीम के और भी साथी बैठे हैं।	बेसबॉल खेल खेला जा रहा है। बल्लेबाज ने लाल जर्सी पहना हुआ है। उसके पीछे दो लोग खड़े हैं। डगआउट में टीम के और भी साथी बैठे हैं।

Figure 3: A sample paragraph for the given image from the dataset created in Hindi

State-of-the-art/baselines	Language modeling	CIDEr	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE-L	METEOR
Top Down Attention (Proposed Method)	LSTM	18.783	6.766	11.688	19.974	35.792	27.490	26.71
Top Down Attention (Baseline-1)	Bi- LSTM	12.666	5.220	9.471	16.930	31.063	27.545	23.684
Adaptive Attention (Baseline-2)	LSTM	17.111	5.838	10.648	19.017	34.948	27.292	26.92
Adaptive Attention MO (Baseline-3)	LSTM	20.74	5.954	10.895	19.339	34.91	27.149	26.426

Table 1: Obtained score with proposed method and baselines

State-of-the-arts/Baselines	CIDEr	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE-L	METEOR
Baseline-1	2.24261e-109	6.40514e-86	1.21712e-86	7.78625e-98	5.22064e-105	4.00075e-68	1.98766e-102
Baseline -2	5.69975e-89	1.07877e-77	1.9483e-79	1.66527e-91	3.27974e-88	1.00585e-52	5.16477e-81
Baseline-3	6.14475e-91	9.60087e-76	3.29491e-87	1.60837e-79	2.40272e-102	1.9483e-79	1.383e-111

Table 2: Welch's t-test based comparison between proposed method and state-of-the-art baselines.

produced captions for the test photographs are pretty accurate and appropriately identify the actions and items in the images.


			
<p>(a) I - एक आदमी एक रसोई में एक रसोई के अंदर खड़ा है आदमी के सामने एक सफेद है । आदमी के पास एक सफेद शर्ट है । महिला के पास एक सफेद शर्ट है । आदमी एक सफेद रंग की शर्ट पहने हुए है । वह एक सफेद रेफ्रिजरेटर के सामने खड़ा है । एक सफेद दीवार के सामने बैठा है । कमरे के सामने एक दीवार है । दीवार पर एक सफेद है ।</p> <p>II- One man one kitchen in one kitchen standing inside man of in front white. Man of has one white shirt. Women of one white shirt. Man a white shirt wearing. He one standing in front of white refrigerator. He white wall in front of seating. Room in front one wall is. Wall of one white is.</p> <p>III- Ek adami ek rasoi me ek rasoi ke andar khara hai adami ke samne ek safed hai. Adami ke pas ek safed shirt hai. Mahila ke pas ek safed shirt hai . Adami ek safed rang ki shirt pahne hue hai. Vah ek safed refrigerator ke samne khara hai. Vah safed deewar ke samne baitha hai. Kamre ke samne ek deewar hai. Deewar par ek safed hai.</p>	<p>(b) I - एक महिला टेनिस कोर्ट पर खड़ी है । उसने एक सफेद शर्ट और सफेद शॉर्ट्स पहने हुए है । महिला ने एक सफेद टैंक टॉप और सफेद स्कर्ट पहन रखी है । वह एक सफेद रैकेट पकड़े हुए है । महिला के पीछे एक सफेद है । लड़की के पीछे एक बड़ा सफेद है । अदालत के पीछे एक आदमी खड़ा है । एक सफेद रंग की टेनिस कोर्ट है ।</p> <p>II- One women tennis court on standing. She one white shirt and white shorts wearing. Women of one white top and white skirt wearing. Women of behind one white. Girl behind one big white is. Court of behind one man standing. One white color of tennis court is.</p> <p>III- Ek mahila tennis court par khari hai. Usne ek safed shirt and safed shorts wearing. Mahila ne ek safed tak top aur safed racket pakre hue hai. Mahila ke piche ek safed hai. Ladaki ke piche ek bara safed hai. Adalat ke piche ek adami khara hai. Ek safed rang ki tennis court hai.</p>	<p>(C) I - एक इमारत के सामने एक सड़क है । भवन के सामने एक पोल है । इमारत के सामने सड़क पर एक पोल है । सड़क के किनारे एक पोल पर एक सफेद है । गली के सामने एक सफेद पोल है । भवन के सामने सड़क के एक सफेद कार है । एक पोल के सामने फुटपाथ पर एक काला पोल है । पोल के सामने एक इमारत है ।</p> <p>II- One building of in front of one road is. Building in front of one poll is. Building in front of road on one poll is. Road of side one poll on one white is . Lane in front of one white poll is. Building in front of road on one white car is. One in front of footpath on one black poll is. Poll in front of one building is.</p> <p>III - Ek imarat ke samne ek sadak hai. Bhawan ke samne ek poll hai. Imarat ke samne sadak par ek poll hai. Sadak ke kinare ek poll par ek safed hai. Gali ke kinare ek safed poll hai. Bhawan ke samne sadak ke ek safed kar hai. Ek ke samne footpath par ek kala poll hai. Poll ke samne ek imarat hai.</p>	<p>(d) I - एक सड़क के किनारे एक स्टॉप साइन है । स्टॉप साइन के सामने एक सफेद है । संकेत के सामने एक स्टॉप साइन है । सड़क के बगल में एक सफेद ट्रक है । सड़क के सामने सड़क पर एक सफेद कार है । गली के सामने सड़क के एक सफेद वैन है । इमारत के सामने एक सड़क है । एक के पीछे एक सफेद इमारत है ।</p> <p>II- One road of side one stop sign is. Stop sign in front of white is. Indication in front of one stop sign is. Beside road one white truck is. In front of road on road one white van is. Building in front of one white is. One of behind one white building is.</p> <p>III- Ek sadak ke kinare ek stop sign hai. Stop sign ke samne ek safed hai. Sanket ke samne ek stop sign hai. Sadak ke bagal me ek safed truck hai. Sadak ke samne sadak par ek safed car hai. Gali ke samne sadak ke ek safed van hai. Imarat ke samne ek sadak hai. Ek ke piche ek safed imarat hai.</p>

Figure 4: Generated paragraph by the proposed method on test images. Here, I, II and III denote the Hindi generated caption, gloss annotation and transliteration, respectively.

5.3 Quantitative Analysis

Although the qualitative analysis has been carried out manually, to conduct the quantitative analysis, a subjective score is still required. The generated Hindi paragraphs here are evaluated against the ground truth paragraph. We perform the qualitative analysis using the BLEU score (Papineni et al., 2002); using n-grams, METEOR (Denkowski and Lavie, 2014), and CIDEr (Vedantam et al., 2015) scores.

We validate our proposed approach and compared it to different baselines using BLEU, CIDEr, and METEOR scores, as shown in Table 1. The results show that our proposed approach outperforms all current baselines.

5.4 Statistical Significance Test

We conduct a statistical significance test (Welch, 1947) at a 5% (0.05) significance level to ensure that the performance increase achieved by our technique

is statistically significant. This test provides the p-values; the lower the p-values, the greater the significance compared to state-of-the-art approaches. We obtain all of the values less than 0.05 (as shown in Table 2), establishing the statistical significance of our technique and demonstrating that the improvement gained by the proposed technique is not by coincidence.

6 Conclusion and Future Works

We present a novel framework for generating paragraphs from photographs in Hindi, which incorporates a region proposal network-based convolutional neural network and an LSTM-based encoder-decoder model with attention mechanisms. We analyze various encoder-decoder models to find the best architecture for paragraph generation from images in Hindi. This work could be extended further by using a transformer-based architecture for language modeling.

Acknowledgment Dr. Sriparna Saha would like to acknowledge the support received from the Young Faculty Research Fellowship program of Visvesvaraya Ph.D. Scheme of Ministry of Electronics Information Technology, Government of India, being implemented by Digital India Corporation (Formerly Media Lab Asia) for conducting this research.

References

- [Anderson et al.2018] 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)*, June.
- [Cornia et al.2020] 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)*, June.
- [Denkowski and Lavie2014] Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.
- [Farhadi et al.2010] 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.
- [Feng et al.2019] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Unsupervised image captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.
- [Goodfellow et al.2013] Ian Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. 2013. Maxout networks. In *International conference on machine learning*, pages 1319–1327. PMLR.
- [Hochreiter and Schmidhuber1997] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- [Hodosh et al.2013] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.
- [Jiang et al.2018] Wenhao Jiang, Lin Ma, Xinpeng Chen, Hanwang Zhang, and Wei Liu. 2018. Learning to guide decoding for image captioning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- [Johnson et al.2016] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Denscap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574.
- [Kingma and Ba2014] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- [Krause et al.2017] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–325.
- [Krishna et al.2016] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*.
- [Liang et al.2017] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. 2017. Recurrent topic-transition gan for visual paragraph generation. In *Proceedings of the IEEE international conference on computer vision*, pages 3362–3371.
- [Lin et al.2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- [Liu et al.2020] 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.
- [Lu et al.2017] 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.
- [MacLeod et al.2017] 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.
- [Mishra et al.2021a] 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.
- [Mishra et al.2021b] Santosh Kumar Mishra, Rijul Dhir, Sriparna Saha, Pushpak Bhattacharyya, and Amit Ku-

- mar Singh. 2021b. Image captioning in hindi language using transformer networks. *Computers & Electrical Engineering*, 92:107114.
- [Papineni et al.2002] 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.
- [Ren et al.2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.
- [Rennie et al.2017] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.
- [Vedantam et al.2015] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.
- [Vinyals et al.2015] 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.
- [Vinyals et al.2016] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663.
- [Welch1947] Bernard L Welch. 1947. The generalization of student’s problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.
- [Xu et al.2015] 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.
- [Young et al.2014] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.
- [Yu et al.2016] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593.
- [Zhou et al.2020] 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.