Encoder-Decoder Based Image Caption Generation Framework for Assamese

Ringki Das[1] and **Thoudam Doren Singh**[1]

[1]Department of Computer Science and Engineering, NIT Silchar, India
{ringkidas,thoudam.doren}@gmail.com

Abstract

Caption generation is an artificial intelligence problem that straddles the line between computer vision and natural language processing. Although significant works have been reported in image captioning, the contribution is limited to English and few major languages with sufficient resources. But, no work on image captioning has been reported in a resource-constrained language like Assamese. With this inspiration, we propose an encoder-decoder based framework for image caption generation in the Assamese news domain. The VGG-16 pre-trained model at the encoder side and LSTM with an attention mechanism are employed at the decoder side to generate the Assamese caption. We train the proposed model on the dataset built in-house consisting of 10,000 images with a single caption for each image. We explain the experimental results in terms of quantitative and qualitative outcomes that support the usefulness of the caption generation model. The proposed model shows a BLEU score of 12.1 outperforming the baseline model.

1 Introduction

Image caption generation is a new and exciting topic in artificial intelligence that has sparked much interest and has been studied extensively in recent years. To interpret the visual contents, computer vision and natural language processing are necessary. As a result, both semantic and linguistic information about the image is required. Expressing the semantic content like human and grammatically correct is a challenging task. Caption generation has a wide range of potential real-world applications. It is helpful for visually impaired people to understand the content of the image. It can also be employed in self-driving automobiles and image search engines. This puts a halt to a slew of important real-world applications, prompting researchers to develop a better model for generating captions in the same way that humans do. Journalists can use news image captioning to describe the contents of the news as well as multimedia analytics.

Image captioning requires recognizing the important objects, attributes, and relationships in an image to generate syntactic and semantically correct description. Earlier caption generation works are based on template and information retrieval. The template-based approach generates the captions by extracting the actions, objects and other attributes in an image and filling them into a pre-defined template. In comparison, the information retrieval-based approach needs a large image database that extracts the visually similar image and generates an image caption by using the caption of the retrieved image. Nowadays, most models are based on deep learning architecture (Bai and An, 2018). The study found that the caption generated using the deep learning approach is more expressive and fluent than the traditional caption generation approaches.

Several significant image caption generation works are proposed in English using deep learning approach. According to our study, image caption generation in Assamese language is still at infancy stage. The Assamese language belongs to the Indo-European language family and it is spoken mainly in the state of Assam in India by approximately 15 million people. In this paper, we propose a caption generation model based on encoder-decoder architecture on news images. At the encoder side, a VGG-16 pre-trained model is used to represent the visual features of an image and to generate the Assamese caption, an LSTM layer with an attention mechanism is employed at the decoder side.

A large set of images and good-quality captions are required for a caption generation system. Existing English datasets are Flickr8K (Hodosh et al., 2013), Flickr30k (Plummer et al., 2015), MS-COCO (Lin et al., 2014), Lifelog (Dang-Nguyen

Proceedings of the 18th International Conference on Natural Language Processing, pages 231–239
Silchar, India. December 16 - 19, 2021.

et al., 2017), Visual Genome (Krishna et al., 2017), Multi30k (Elliott et al., 2016) etc. However, there is currently no comparable annotated dataset accessible in Assamese. Dataset scarcity is a major challenge, particularly for a morphologically rich (Saharia et al., 2012) language like Assamese. Therefore, we present an annotated dataset for caption generation in the news domain and forward it for future research. First, we collect the news articles consisting of both images and text from the local Assamese newspapers. We pre-process the data as an initial step after the collection. After that, each news image is manually annotated with one description. Furthermore, we evaluate the model using both quantitative and qualitative parameters. The proposed system is one of the earliest reported Assamese news image caption generation model and the experiment results of the developed model are promising. The objectives of this paper are:

1. The goal at hand is to build a news image caption generation model for the Assamese language in a low-resource scenario. An attention mechanism-based model is compared to the baseline model. The model is evaluated against predefined metrics to describe the news image.

2. We trained the proposed model using an in-house built dataset containing 10,000 images with single caption for each image.

2 Related Work

Fang et al. (2015) suggested a caption generation system on the MS-COCO dataset, which received a BLEU-4 score of 29.1. A visual recurrent representation model was suggested by Chen and Lawrence Zitnick (2015) for image caption generation on the MS-COCO, Flickr 8K and 30K and PASCAL 1K datasets. To describe and visualize the image caption, a bi-directional mapping between images and sentence-based descriptions was carried out. Karpathy and Fei-Fei (2015) also described the image region using a multimodal recurrent neural network on Flickr8K, Flickr30K and MS-COCO datasets. A Chinese image caption generation model was introduced by Peng and Li (2016) on Flickr30k and MS COCO dataset. They demonstrated that a character-level strategy is more effective than a word-level one. Soh (2016) reported a top-down caption generation strategy employing CNN-LSTM architecture on the MS COCO dataset with a 3.3 BLEU score. Miyazaki and Shimizu (2016) proposed a deep recurrent network based image caption generation model on the cross-lingual domain. The YJ Captions 26k Dataset, a Japanese version of the MS-COCO dataset, was built for this purpose. Amritkar and Jabade (2018) reported an image caption generation with CNN and RNN architecture on Flickr8k and MS COCO datasets.

An attention-based remote sensing image captioning system was reported by Lu et al. (2017) on their own built remote sensing caption generation dataset. They employed both hard and soft attention mechanisms to train the model. They found that the hard attention mechanism performed better than the soft attention mechanism. Dhir et al. (2019) used attention-based architecture to report a Hindi caption generation approach. They manually translated the MS COCO dataset into Hindi for the dataset. You et al. (2016) developed a semantic image attention model to concentrate on the linguistically significant image object.

Batra et al. (2018) proposed an encoder-decoder-based news image caption generating architecture on the BBC news data. The model takes an image from the news related to news documents as input and outputs an appropriate image caption.

Rahman et al. (2019) introduced Chittron, a Bangla image captioning model. A total of 16,000 images was collected and has been manually annotated by two native Bangla speakers. Next, a VGG-16 image embedding model integrated with a stack LSTM layer is used to train the model. The proposed model has gained a BLEU score of 2.5. Again Kamal et al. (2020) used deep learning techniques to create an automated Bangla caption generation system called TextMage on the BanglaLekhaImageCaptions dataset. The TextMage model could understand the visual scenes that belong to the Bangladeshi geographical context. The proposed model is trained on the BanglaLekhaImageCaptions dataset consisting of 9,154 images along with two descriptions for each image. The use of Visual Genome image captioning in a multimodal machine translation challenge was reported by Meetei et al. (2019b). The generation and evaluation of Hindi image caption on the Visual Genome dataset were carried out by Singh et al. (2021a). Additionally, attention based image and video caption generation framework were carried out by Singh et al. (2021c), Singh et al.

Table 1: Statistics of the dataset

Data set	Image	Caption
Train	8000	8000
Development	1000	1000
Testing	1000	1000
Total	10000	10000

(2021b). Meetei et al. (2019a) reported a work on identifying the Manipuri and Mizo texts in an image that is a crucial challenge in image captioning.

3 An Assamese Multimodal Dataset

Several benchmark datasets for caption generation are Flickr8K, Flickr30K and MSCOCO, available in English, but none are accessible for resource-constrained languages, including Assamese. So, we built an Assamese multimodal dataset from the news domain. We carried out the following dataset preparation steps:

1. Collection of data
2. Image pre-processing
3. Image annotation

3.1 Data collection

The preparation of a standard dataset is one of the most challenging parts of a deep neural network model. A newspaper has both images and text, making it a valuable source of information. To address the data set availability problem, we collected 10,000 Assamese news images from three Assamese local e-newspapers, namely Ganaadhikar[1], Niyomia Barta[2] and Asomia Pratidin[3]. The data is collected during June 2020 and April 2021. After pre-processing and annotation, the raw information is used to train our model. Each news image is manually annotated with an insightful narrative relevant to the news event by two native Assamese speakers. A statistics of Assamese news caption dataset is presented in Table 1.

3.2 Image pre-processing

Based on the news event, the original image is manually cropped to highlight the essential portion of the image to extract the relevant part of the image. The specific object features of images must be combined in the correct order to correlate with the caption. A sample data is shown in Figure 1 where the news is about the baby; therefore, we crop the image wherein the Figure 1B focus on the baby only.

3.3 Image annotation

Describing the image content is one of the important tasks of a caption generation model. Each image has been manually annotated with one image caption from the news content by two native Assamese speakers. At times, the content and the image cannot convey the same meaning. As a result, the annotators have labeled each image with a more appropriate news caption to the event as part of the post-editing process. Some news articles have only the logo or file images which are not relevant for the image. These irrelevant images are filtered out. Then, we put a correct news description by performing a manual post-editing of the captions for a better captioning model. Figure 2 shows one sample of the Assamese news caption dataset.

4 System Architecture

The first stage of an image caption generation model is image feature extraction, and the second part is image description generation. A convolutional neural network is used to extract the image features at the encoder side and LSTM layers are used to train the language model for image description at the decoder side. This paper describes the development of an encoder-decoder based image caption generation framework using CNN-LSTM architecture with attention mechanism in the Assamese language news domain. The proposed model consists of three phases:

1. Text pre-processing
2. Image feature extraction
3. Caption generation

4.1 Text pre-processing

Before feeding the text input to the neural network, it is important to pre-process the text data and transform it into a numerical form. For input text representation, a word embedding layer is used. It provides a dense representation of the input text and then passes it to the next LSTM layer.

[1]http://ganaadhikar.com
[2]https://niyomiyabarta.org/home/
[3]https://www.asomiyapratidin.in/

Figure 1: Image cropping

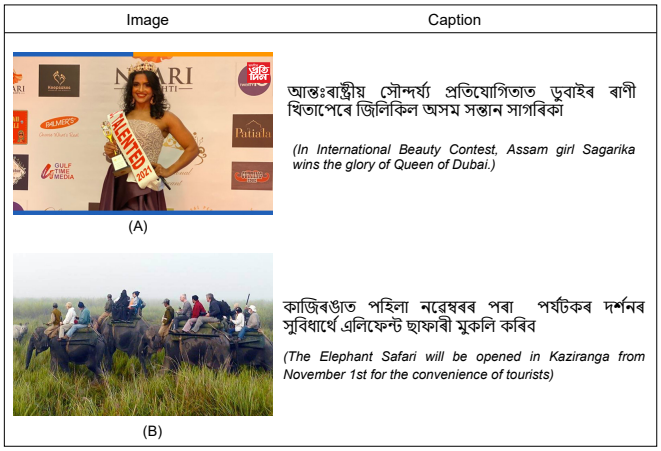

Figure 2: An example of Assamese caption dataset

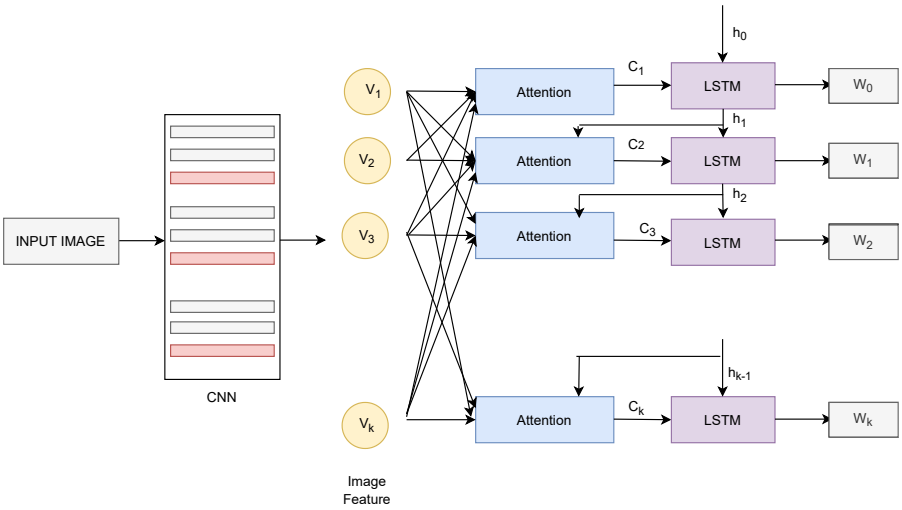

Figure 3: A graphical representation of proposed architecture

4.2 Image feature extraction

A convolutional neural network (CNN) is deployed to extract the image features. It encodes an image into an intermediate vector representation. To extract a feature set vectors, CNN is employed as an encoder. In this framework, we use VGG-16 (Simonyan and Zisserman, 2014) pre-trained CNN model as an encoder. VGG-16 model is trained on the ImageNet [4] dataset. It encodes the input image into a fixed-length vector for further processing to generate the image description. For image

[4]https://image-net.org/

feature extraction, the input image is resized into 224 × 224 dimensions. We discard the output of the last layer and stored the output of 6_{th} layer. The dimension of each feature vector is 7× 7 × 512.

4.3 Caption generation

To solve the vanishing gradient problem, the long short-term memory (Hochreiter and Schmidhuber, 1997) is employed as a decoder that can learn long-term dependencies. It is used for language modeling trained on text data to predict the next word. LSTM is trained so that it can produce the caption by generating one word for every time step conditioned on a context vector, the previous hidden state and the previously generated words. First, the captions are tokenized to create a lexicon of unique words. The size of our vocabulary is 8558. The model understands the start and end of each caption since each sentence is concatenated with the "start" and "end" tags.

4.4 Attention Mechanism

The fixed-length context vector in a sequence to sequence (seq2seq) model fails to remember long sentences. As a result, an attention mechanism (Xu et al., 2015) is utilized to solve this problem. The attention mechanism works on the relevant parts of the input image and ignores the rest. Rather than compressing an entire image into a static representation, attention allows the salient features to dynamically come to the forefront as needed. In simple terms, the context vector is a dynamic representation of the relevant part of the image input at time t. The attention mechanism considers the relevant part of the image when the LSTM generates a new word, so the decoder only uses that part of the image. An attention mechanism is classified into local and global attention mechanisms. Global attention is defined as paying attention to all source parts of an image (Luong et al., 2015). Local attention focuses to only a few source positions (Bahdanau et al., 2014).

In this current work, we employ a global attention mechanism that is placed in all source positions. In between CNN and LSTM, we use the attention mechanism to help the decoder to focus the important parts of the image. The global attention mechanism considers all the hidden states of the encoder while deriving the context vector c_t. In order to compute the context vector c_t, we first compute the variable-length alignment vector a_t. The variable-length alignment vector a_t whose size equals the number of time steps on the source side is derived by comparing the current target hidden state. The encoder hidden states and their respective alignment scores are multiplied to calcuate the context vector. The formula for calculating the context vector, alignment vector and score are listed in equations 1, 2, 3 and 4, respectively.

$$c_t = \bar{h}_s * a_t(s) \tag{1}$$

In equation 1, global context vector c_t is computed as the weighted average of the encoder hidden states $\bar{h}_s$ and alignment vector a_t.

$$a_t(s) = align(h_t, \bar{h}_s) \tag{2}$$

$$= \frac{exp(score(h_t, \bar{h}_s))}{\sum_s exp(score(h_t, \bar{h}_s))} \tag{3}$$

From the equations 2 and 3, the variable-length alignment vector a_t is derived by computing the similarity between current target hidden state h_t with each source hidden state $\bar{h}_s$.

$$score(h_t, \bar{h}_s) = \begin{cases} h_t^T \bar{h}_s & dot \\ h_t^T w_a \bar{h}_s & concatenate \\ v_a^T \tanh(w_a(h_t, \bar{h}_s)) & general \end{cases} \tag{4}$$

Again in equation 4, score is referred as a content-based function for which we consider three different alternatives like dot, concatenate and general, respectively.

4.5 Beam Search

Beam search is an optimization search strategy for reducing memory requirements. The search tree is built using a breadth-first search approach. We use the beam search method to evaluate the captions. For generating sentences of size t + 1, the beam search approach inspects the top k sentences and holds the highest probability one until it reaches the "end" tag or the maximum length of the caption.

Table 2: Adequacy and fluency rating scale

Rating	Adequacy	Fluency
5	All	Flawless
4	Most	Good
3	Much	Non-native
2	Little	Disfluent
1	None	Incomprehensive

Sl No.	Reference Caption	Generated Caption	Adequacy	Fluency
1	অসমবাসীলৈ ৰঙালী বিহুৰ শুভকামনা জনালে আমেৰিকাৰ ৰাষ্ট্রপতি জ' বাইডেনে *(US President J. Biden wishes Rangali Bihu to the people of Assam)*	*Proposed system* অসমবাসীলৈবিহুৰশুভকামনাজনালে আমেৰিকাৰ ৰাষ্ট্রপতি জ' বাইডেনে *(US President J. Biden wishes happy bihu to the people of Assam)*	4	5
		Baseline system অসমবাসীলৈ বিহুৰ জনালে ৰাষ্ট্রপতি জ' বাইডেনে *(President J. Biden wishes bihu to the people of Assam)*	3	2
2	কৰোনা ভাইৰাছৰ সংক্রমণ বাবে দেশজুৰি লকডাউন ঘোষণা কৰা হৈছে *(A nationwide lockdown has been announced for the corona virus)*	*Proposed system* ভাইৰাছৰ সংক্রমণ বাবে লকডাউন ঘোষণা কৰা হৈছে *(A lockdown has been announced for the virus)*	4	4
		Baseline system ভাইৰাছৰ সংক্রমণ বাবে ঘোষণা কৰা হৈছে *(A has been announced for the virus)*	3	3
3	নগাঁৱৰ বঢ়মপুৰৰ দুৰ্গম পাহাৰত শোকাৱহ ঘটনা *(Tragedy in the remote hills of Barampur in Nagaon)*	*Proposed system* পৰা সংঘটিত হয় এই ঘটনা *(From occur this incident)*	1	1
		Baseline system এই ঘটনা This incident	1	1

Figure 4: Rating score based on adequacy and fluency

4.6 Training Details

The input of the VGG-16 convolutional neural network is 224x224 RGB images, which produces a vector of size 49x512 for each image. Again the captions are fed into the word embedding layers with 256 neurons. We train our model with 0.5 dropout rate, softmax cross-entropy loss function and Adam optimizer (Kingma and Ba, 2014) with batch size of 64 for 25 epochs. For training, we use 8000 images, and for development and testing, we use 1000 images each. The experimental results demonstrate that the LSTM with attention mechanism as a middle layer showed more effectiveness in generating the image captions.

5 Experimental Results and Discussion

The quantitative and qualitative analysis of the proposed system is covered in this section.

5.1 Quantitative analysis

For quantitative analysis, the BLEU (Papineni et al., 2002) metric is used. It checks the similarity of a generated output sentence corresponding to a reference sentence. We report the BLEU scores of the baseline and proposed models in Table 3. The formula for calculating the BLEU score is listed in equations 5 and 6.

$$BP = \begin{cases} 1 & \text{if } c>r \\ 0 & \text{if } c<=r \end{cases} \quad (5)$$

$$BLEU = BP * exp(\sum_{n}^{N} w_n \log P_n) \quad (6)$$

where,
c is candidate sentence length
r is reference sentence length
P_n is n-gram precision
w_n is weight

5.2 Qualitative Analysis

To verify the correctness of the machine-generated output, two native speakers of Assamese evaluate the generated captions by using the criterion set by linguistic data consortium(LDC) (Denkowski and Lavie, 2010). A sample example based on adequacy and fluency rating scale is shown in Figure 4 and it is found that most of the captions are flawless. According to LDC, human judgment is classified into adequacy and fluency categories (Table 2).In comparison to the source text, adequacy refers to how much meaning the target text can express. Fluency is the capability to describe a grammatically correct target text. To calculate the adequacy and fluency score, we randomly pick 100 sentences

Table 3: BLEU score of our proposed and baseline architectures

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Baseline	31.8	25.3	20.5	11.4
Proposed	40.4	31.6	21.8	12.1

News image	Reference caption	Generated caption
 (A)	নৰেন্দ্ৰ মোদী আজি দেশৰ উদ্দেশ্যে ভাষণ দিছে *(Narendra Modi is addressing the nation today)*	*Proposed system* নৰেন্দ্ৰ মোদীৰ সংবাদমেল *(Conference of Narendra Modi)* *Baseline system* মোদীৰ বৈঠক *(Conference of Modi)*
 (B)	অগ্নিগৰ্ভা নাগালেণ্ডত আৰক্ষী নিহতৰ সংখ্যা ১৪জনলৈ বৃদ্ধি *(Firefight death toll rises to 14 in Nagaland)*	*Proposed system* ভয়াবহ অগ্নিকাণ্ড আৰক্ষী নিহত *(Horrific fire kills police)* *Baseline system* ভয়াবহ অগ্নিকাণ্ড *(Horrific fire)*
 (C)	টীম ইণ্ডিয়াৰ তাৰকা বিৰাট কোহলি কঠোৰ অনুশীলনত মগ্ন হৈ আছে। *(Team India's star Virat Kohli is immersed in rigorous training.)*	*Proposed system* অনুশীলনত মগ্ন কোহলি *(Kohli immersed in practice)* *Baseline system* বিৰাট কোহলি *(Virat Kohli)*
 (D)	ভাৰতত কৰোনাত আক্ৰান্ত হৈ মৃতুৎ হোৱা লোকৰ সংখ্যা বৃদ্ধি *(In India, the number of death troll due to corona is increasing)*	*Proposed system* ভাৰতত কৰোনাত মৃতুৎ লোক বৃদ্ধি *(Death toll due to corona rises in India)* *Baseline system* কৰোনা *(corona)*

Figure 5: Generated captions by the proposed and baseline models

Table 4: Human evaluation results

Model	Adequacy	Fluency
Baseline	1.48	1.96
Proposed	1.91	2.05

from the test dataset. Table 4 shows the scores for adequacy and fluency respectively.

5.3 System Comparison

Till today, no model has been reported in Assamese news image captioning to the best of our knowledge. To make a fair comparison, we propose a baseline model of CNN-LSTM architecture (Vinyals et al., 2015) and compared with the proposed model as shown by Table 3.

5.4 Discussion

Figure 5 shows the sample input and output for the image caption generating model. From the Figure 5A, the model can detect the named entity, i.e., "Narendra Modi" and also generate the caption about conference, which is a good result. The caption that is generated is meaningful, although it is less fluent. Therefore, the adequacy and fluency are considered as 4 and 3, respectively, from the point scale rating (Table 2). As shown in Figure 5B, the model can also show a good result. The model can detect the "fire" and the "army" as part of the image. The machine-generated caption can convey the meaning. In this example, the adequacy and

fluency scores are 5 and 4, respectively. As shown in Figure 5C, the model identifies the named entity, i.e., "Virat Kohli" and also generated caption says about the action. Thus, the generated caption can convey the meaning and is fluent. As a result, both adequacy and fluency receive a 5 on the point scale. After seeing the "PPE kit", costume, the model can generate about the "corona" in Figure 5D. But the generated caption is not fluent. So the adequacy and fluency rating is 3 and 2, respectively.

5.5 Error analysis

There are couple of reason why the generated captions are imperfect. Poor caption quality can be a major reason for erroneous caption generation. Some image captions and news images are the least connected, which is unusual. As a result, some articles contain merely logo images or image files unrelated to current events. The absence of specific functional tokens in the training caption is another reason for poor quality generated caption.

6 Conclusion and Future Work

In this paper, we report a CNN-LSTM based framework with an attention mechanism for Assamese caption generation on the multimodal news dataset. The attention mechanism decides where to pay attention in order to generate a meaningful caption. We also report the findings of the investigation of caption generation on the Assamese language on low resource setting. To assess model performance, we used both qualitative and quantitative approaches. It is observed that the proposed framework outperforms the baseline model. In future, various architectures such as ResNet with mBERT or Indic BERT may be explored for any significant improvement of the system results further. We intend to expand the dataset in the future with a more diverse and wide collection of images of various domain-specific events, each with several appropriate descriptions, in order to build a human-like caption.

Acknowledgment

We acknowledge CNLP (Centre for Natural Language Processing) at NIT Silchar for giving access to the lab.

References

Chetan Amritkar and Vaishali Jabade. 2018. Image caption generation using deep learning technique. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pages 1–4. IEEE.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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

Vishwash Batra, Yulan He, and George Vogiatzis. 2018. Neural caption generation for news images. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Xinlei Chen and C Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2422–2431.

Duc-Tien Dang-Nguyen, Liting Zhou, Rashmi Gupta, Michael Riegler, and Cathal Gurrin. 2017. Building a disclosed lifelog dataset: Challenges, principles and processes. In *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, CBMI '17, New York, NY, USA. Association for Computing Machinery.

Michael Denkowski and Alon Lavie. 2010. Choosing the right evaluation for machine translation: an examination of annotator and automatic metric performance on human judgment tasks.

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, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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.

Abrar Hasin Kamal, Md Asifuzzaman Jishan, and Nafees Mansoor. 2020. Textmage: The automated bangla caption generator based on deep learning.

In *2020 International Conference on Decision Aid Sciences and Application (DASA)*, pages 822–826. IEEE.

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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

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.

Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. 2017. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2019a. Extraction and identification of manipuri and mizo texts from scene and document images. In *International Conference on Pattern Recognition and Machine Intelligence*, pages 405–414. Springer.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2019b. Wat2019: English-hindi translation on hindi visual genome dataset. In *Proceedings of the 6th Workshop on Asian Translation*, pages 181–188.

Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1780–1790.

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 of the Association for Computational Linguistics*, pages 311–318.

Hao Peng and Nianhen Li. 2016. Generating chinese captions for flickr30k images.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

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

Navanath Saharia, Utpal Sharma, and Jugal Kalita. 2012. Analysis and evaluation of stemming algorithms: a case study with assamese. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, pages 842–846.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Alok Singh, Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2021a. Generation and evaluation of hindi image captions of visual genome. In *Proceedings of the International Conference on Computing and Communication Systems: I3CS 2020, NEHU, Shillong, India*, pages 65–73. Springer.

Alok Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2021b. Attention based video captioning framework for hindi. *Multimedia Systems*, pages 1–13.

Alok Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2021c. An encoder-decoder based framework for hindi image caption generation. *Multimedia Tools and Applications*, pages 1–20.

Moses Soh. 2016. Learning cnn-lstm architectures for image caption generation. *Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep*.

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.

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. PMLR.

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.