Multimedia Tools and Applications (2022) 81:10051–10069
https://doi.org/10.1007/s11042-022-12042-8

Assamese news image caption generation using attention mechanism

Ringki Das[1] · Thoudam Doren Singh[1]

Received: 6 January 2021 / Revised: 4 June 2021 / Accepted: 4 January 2022/
Published online: 14 February 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

In recent times, neural networks and deep learning have made significant contributions in various research domains. In the present work, we report automatic caption generation of an image using these techniques. Automatic image caption generation is an artificial intelligence problem that receives attention from both computer vision and natural language processing researchers. Most of the caption generation tasks exist in the English language and no work has been reported yet in Assamese to the best of our knowledge. Assamese is an Indo-European language spoken by 14 million speakers in the North-East region of India. This paper reports the image caption generation on the Assamese news domain. A quality image captioning system requires an annotated training corpus. However, there is no such standard dataset available for this resource-constrained language. Therefore, we built a dataset of 13000 images collected from various online local Assamese e-newspapers. We employ two different architectures for generating the news image caption. The first model is based on CNN-LSTM and the second model is based on the attention mechanism. These models are evaluated both qualitatively and quantitatively. Qualitative analysis of the generated captions is carried out in terms of fluency and adequacy scores based on a standard rating scale. The quantitative result is evaluated using the BLEU and CIDEr evaluation metrics. We observe that the attention mechanism-based model outperforms the CNN-LSTM based model for our task.

Keywords News caption generation · CNN-LSTM architecture · Attention mechanism · Assamese news · Resource-constrained language

1 Introduction

An automatic generation of natural language description for an image is known as image caption generation. The image captioning task is more complex than object recognition. It

✉ Ringki Das
ringkidas@gmail.com

Thoudam Doren Singh
doren@cse.nits.ac.in

[1] Department of Computer Science and Engineering, National Institute of Technology, Silchar, India

is an emerging and fascinating topic in natural language processing that has seen an upsurge of research and is widely investigated for the last few years. Earlier image captioning work was mainly information retrieval and template-based. But, nowadays most of the models are based on deep neural network architecture [3]. The recent progress improves the system performance though the results are not sufficiently satisfying. There is a huge gap between machines and the human brain in terms of understanding an image and generating its description [15]. A machine can not work simultaneously like humans because devices are precise but can not cover all the scenarios. As a result, caption generation will remain a challenge to be solved among the researchers.

Describing an image with just a glance is an easy task for a human. But, a machine can not explain and understand the visual scene easily. The quite challenging part of image captioning is the semantic content of the image and expressing it in an organized human-like and grammatically correct sentence. Understanding the image contents requires computer vision and natural language processing for the language model. Caption generation has many promising real-world applications in different areas like a guidance device for a visually impaired, visual guided machine translation and multimodal sentiment analysis etc. The model can be used in self-driving cars and image searching tools as well. This pauses so many meaningful real-life applications which trigger researchers to build a better model to generate captions like humans. Apart from this, news image captioning is the most common technique and the biggest challenge that is helpful for journalists for describing the news contents. Thus, news caption generation finds its applications in news, media and multimedia analysis.

In today's civilization, online news gives us up-to-date global news and is an essential part of our daily lives. People can often communicate with the world through natural language. But images are also a beautiful way to communicate and understand our surroundings. As an old saying, 'paint a picture in your mind's eye.' A good image can express things more loudly than a thousand words. Description of the image is a multimodal task that requires both semantic information and linguistic processing of the image. A more significant number of image caption generation work exist in the English language. The previous work on news image caption generation is scarce. To the best of our knowledge, no work on the Assamese language has been proposed in image caption generation. So, there are many unsolved problems to explore for this resource-poor language. Due to the sparsity of annotated resources, caption generation is a challenging task for many languages other than English.

This paper focuses on two seq2seq deep neural network architecture to address the Assamese news image caption generation. News data is an excellent source of multimedia files that consists of both image and natural language text. The proposed news caption generation system has been developed to understand the visual scenes of the newspaper. The first model is CNN-LSTM architecture and the other model is on the attention mechanism. Therefore, in the very first step, we pre-trained the Assamese news dataset using a convolutional neural network. CNN is working as an encoder to extract the image features. In contrast, the LSTM layer works as a decoder for decoding sentences to generate a vector representation of the image caption. The automated news image describing system follows a VGG-16 as a feature extractor from an image. Next, the feature vector is fed to the LSTM network to generate the image caption. Again, in the second model, we deploy an attention mechanism on top of the LSTM layer. The attention mechanism is used to decide where to put attention and what information to analyze and plays an essential role in the next movement. We employ a hard attention mechanism to focus only on the important part of the news image.

The Assamese language belongs to the Indo-European language family. As per the 2011 census, it is spoken mainly in the state of Assam in India by approximately 15 million people. A caption generation system requires a rich dataset with images along with good quality captions. Some of the existing corpora in English are MS-COCO, Flickr8K, Flickr30k, Lifelog, Visual Genome, etc. But, there is no such annotated dataset available in Assamese at present. The scarcity of datasets is a big issue, especially for a morphologically rich language like the Assamese. Therefore, we collect 13000 news images embedded within the article from different Assamese daily newspapers. After that, native speakers of this language manually annotated each news image with one description respectively. The essential objective and commitments of this paper are summarized as follows:

1. In our proposed architecture, we aim to develop a news image caption generation model in low resource Assamese language. Our model is well suited as a baseline for future work on the Assamese news image caption generation.
2. We introduce an Assamese news dataset. Development of an Assamese contextual newspaper dataset consists of both news images embedded with news articles. Next, native speakers of this language manually annotated each image with one image caption related to the news event.
3. We demonstrate the effectiveness of the Assamese news dataset on two different deep neural architectures based on CNN-LSTM and local attention mechanisms respectively.
4. Finally, we evaluated the models against the predefined evaluation metrics for describing the news images. We also conduct a human analysis of results as well as the error analysis.

The remaining part of the paper is organized as related work is discussed in Section 2, the methodology is described in Section 3, the experimental results and analysis are given in Section 4 and the paper is concluded with future direction in Section 5.

2 Related work

The growing accessibility of online data is a blossom for artificial intelligence and other related research areas. The neural network becomes a rampant field day by day, coming out with many applications like machine translation, image captioning, sentiment analysis, etc. According to the prior literature survey, many different techniques have been developed till today's date. Various approaches for caption generation are carried out to overcome the challenges. In this section, we present other existing methods for the caption generation used to develop the system.

2.1 Image caption generation

An automated Bangla caption generator system named TextMage was developed on the BanglaLekhaImageCaptions dataset by Kamal et al. [16] using deep learning techniques. Feng et al. [8] proposed a caption generation framework on the MS-COCO dataset and achieved a BLEU-4 score of 29.1. After analyzing the image, they detect the words and generate captions followed by re-ranking. A convolutional neural network was used to extract features from every region and finally combined the MIL (multiple instance learning) information to select the most suitable caption. A recurrent visual representation-based model was suggested by Chen et al. [6] for image caption generation. A bi-directional mapping was done between images and sentence-based descriptions to describe and visualize the image caption. An image region description model was proposed by Karpathy et al.[17] using a multimodal recurrent neural network. Again, a deep convolutional neural network architecture for image caption generation was reported by Amritkar et al. [1] on the Flickr8k dataset. To generate entity-aware image description, Lu et al. [20] proposed a caption generation model using CNN-LSTM architecture. They used a knowledge-graph algorithm for filling the template with specific named entities and other relevant data like the time of the photo and geo-location. A generative model for caption generation was reported by Haripriya et al. [12] on the MS-COCO dataset using an encoder-decoder architecture. Soh et al. [35] also introduced a top-down image caption generation model using CNN-LSTM architecture. A two-phase Chinese image caption generation on Flickr30k images was proposed by Peng et al. [28]. As there is no space between Chinese words, they tokenized the Chinese sentence into words and fed it into RNN. The next phase split the Chinese sentence into characters and fed it into the same model. It was shown that the character level method works better than the word level method. Describing remote sensing image content is not clear and accurate. Lu et al. [19] proposed a remote sensing caption generation model by using deep encoder-decoder architecture. An automatic Bangla image caption generation system named Chittron was introduced by Rahman et al. [30]. The deep neural network architecture was trained using the VGG-16 and stack LSTM layer. A cross-lingual generative image caption generation model on the deep recurrent network was suggested by Miyazaki et al. [25]. The Japanese version of the MS-COCO dataset named YJ Captions 26k Dataset was developed for this purpose. Xu et al. [38] introduced the attention mechanism for the image caption generation model. Again, an attention-based image caption generation model was proposed by Mansimov et al. [22] on the Microsoft COCO dataset. An attention-based architecture was suggested by Dhir et al. [7] for generating image caption in the Hindi language. For the dataset, they have manually translated the MS COCO dataset into Hindi. A semantic attention model was suggested by You et al. [39] to focus on the linguistically important image object. Meetei et al. [24] reported the use of image captioning of Visual Genome in a multimodal machine translation task. Singh et al. [34] carried out the generation and evaluation of Hindi image captions of Visual Genome. Meetei et al. [23] also reported the extraction of written Manipuri and Mizo texts in an image which is one of the important challenge in image captioning.

2.2 News image caption generation

An encoder-decoder-based news image caption generation architecture is proposed by Batra et al. [4] on the BBC news dataset. The input to the model is a news image associated with documents, and as an output, the model generates an appropriate image caption. Prajapati et al. [29] reported an automatic caption generation framework on a news dataset. They extract the keywords using the LDA (Linear Discriminant Analysis) algorithm. Using the maximum overlapping keywords, the appropriate news image caption is generated. A content-based approach for developing a news caption generation model was introduced by Kohakade et al. [18]. They identified the named entities by using named entity recognition (NER) module and a face recognition model. Feng et al. [9] reported a caption generation model by using a probabilistic image annotation model on news images. Again, one standard two-stage caption generation architecture is adopted by Feng et al. [10] on BBC news. The two-stage architecture consists of content selection and surface realization.

3 Methodology

This section gives a brief description of the proposed methodology consisting of dataset preparation and system architecture.

3.1 Data acquisition

Dataset is a crucial part of generating the Assamese news caption. Therefore, data acquisition is a very salient and diligent process. Creating a comparable dataset is one of the most challenging parts of developing a deep neural network model. There are some available benchmark datasets in English, but no such dataset is available for several resource-constrained languages. Due to the low availability of the Assamese dataset, data collection becomes a long and exhausting process. A newspaper is a rich source of information that consists of both image and text. Nowadays, the e-paper of different languages are freely available. To address the data set availability issue, a collection of 13000 Assamese news images has been manually accumulated from publicly available Assamese e-newspaper, namely Ganaadhikar[1], Niyomia Barta[2] and Pratidin[3]. The collected dataset consists of news images accompanied with text. The data has been collected from June 2019 to April 2020. The collected dataset is used to train our model after performing the pre-processing and annotation steps. Native speakers of this language manually annotated each news image with an informative description related to the news event. Figure 1 shows examples of news images and their captions in the Assamese. It can be seen from the example that the captions related to the news images provide more information than what is depicted in the image only.

Our next step towards developing a caption generation model is data annotation. Data annotation is a labor-intensive as well as expensive task. Different annotators have different point of view. For that reason, the annotation may vary from person to person with a distinct sense of the sentence. Native Assamese speakers have manually annotated each image with one image caption. All the collected images are kept in the JPEG file format. We use the VGG-16 for image feature extraction. The 224*224 pixels size is standard for VGG-16 [11, 33]. Therefore, the input image is cropped and resized to 224*224 pixels for a computationally efficient feature extraction process. During training, we monitor the system performance on the development of the dataset to decide when to save models for testing. We evaluated the system performance on a test dataset consisting of 1000 images and corresponding captions. Table 1 provides the statistics of our corpus.

3.2 Pre-processing of data

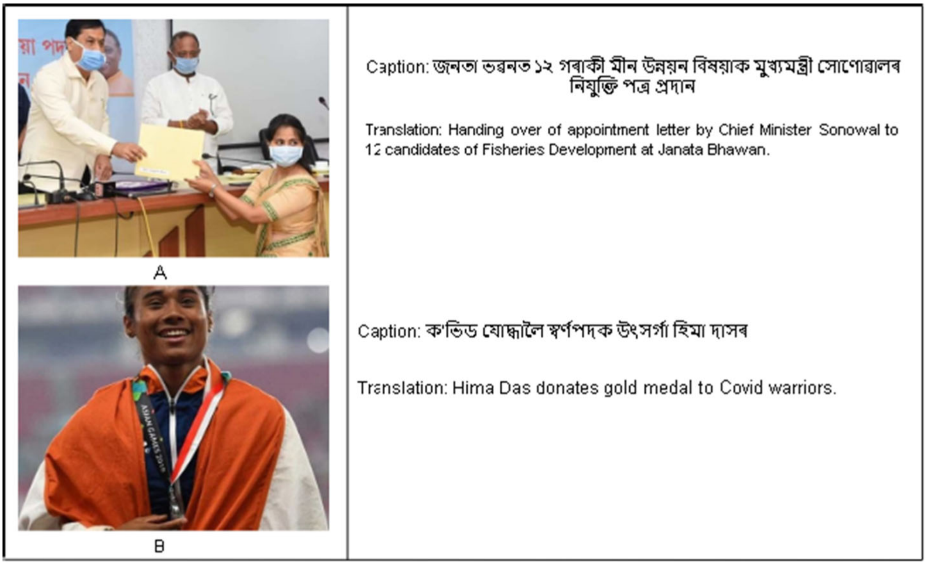
The dataset preparation task begins with data collection followed by a pre-processing task for both images and captions separately. For data pre-processing, image cropping and modifying the image captions are the main steps.

Image Cropping: Sometimes the news images contain lots of objects that are not relevant for the task. Therefore, as part of the pre-processing, the original image is cropped to highlight the important region of the image. The particular object features of images must

[1] http://ganaadhikar.com

[2] https://niyomiyabarta.org/home/

[3] https://www.asomiyapratidin.in/



Caption: জনতা ভৱনত ১২ গৰাকী মীন উন্নয়ন বিষয়াক মুখ্যমন্ত্ৰী সোণোৱালৰ নিযুক্তি পত্ৰ প্ৰদান

Translation: Handing over of appointment letter by Chief Minister Sonowal to 12 candidates of Fisheries Development at Janata Bhawan.

Caption: ক'ভিড যোদ্ধালৈ স্বৰ্ণপদক উৎসৰ্গা হিমা দাসৰ

Translation: Hima Das donates gold medal to Covid warriors.

Fig. 1 Assamese News Caption

be integrated in the correct combinations to relate with the caption. Image cropping has also been performed manually only for some images. A sample is shown in Fig. 2.

Caption Annotation: Caption annotation is the most significant task for the image caption generation model. Native speakers of Assamese have manually annotated each image as a news image caption from the news headline. In a few cases, the headline and the image are unable to convey the same meaning. Then, a single substantive text is prepared from the news description as an image caption. In the MS COCO dataset, there are 20,000 images with five captions per image. But due to different constraints, multiple captions per image are not developed yet in our dataset. A sample example is shown in Fig. 3.

3.3 System architecture

The encoder-decoder architecture is used in the machine translation task for solving sequence to sequence problems. In this architecture, the encoder is used to read a sentence in the source language and convert it into a fixed-length embedding vector. Next, the decoder decodes the embedding vector into the target language. This architecture also achieves great success in the image caption generation. In this paper, we aim to develop a news image

Table 1 Dataset statistics

Data set	Size
Training	11000
Development	1000
Test	1000
Total captions	13000

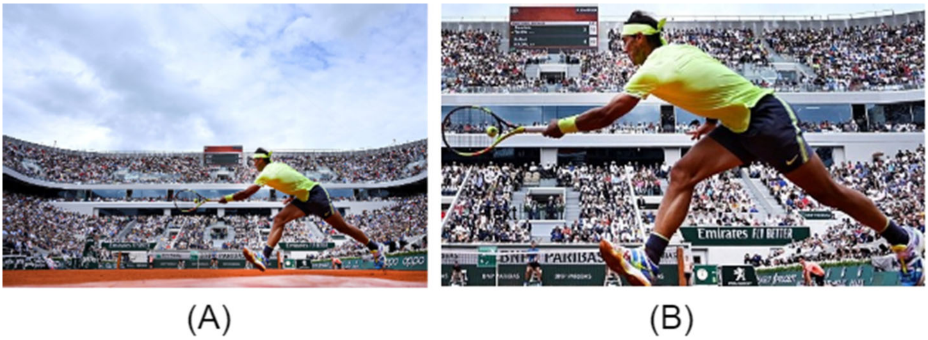


Fig. 2 (A) Original image and (B) Cropped image from (A)

caption generation framework in the Assamese language using CNN-LSTM and attention-based architecture. The formulation of the proposed model is defined as follows: given a news image I, and its accompanying article A, $W = (W_0, W_1 ..... W_n)$ is the sentence description of the news image and $P(W_n | I, W_0, W_1 .... W_n - 1)$ is all the preceding words generated by the model. In the following Sections 3.4 and 3.5, we explain both the models.

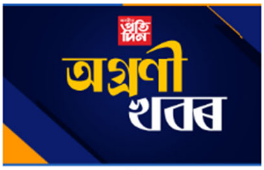
Input Image	Caption
 A	News Caption: চাবুৱাত খেল বিশ্ববিদ্যালয় স্থাপনৰ বাবে অসম চৰকাৰে খৰকতীয় আঁচনি হাতত লৈছে Translation: Assam Government working on war footing to start Sports University in Chabua. Expected Caption: অসমৰ মুখ্যমন্ত্ৰী সৰ্বানন্দ সোণোৱালে ৰাজহুৱা স্থানত লন টেনিছ্ খেলিছে Translation: Assam Chief Minister Sarbananda Sonowal playing lawn tenning in public place.
 B	News Caption: জিলা পৰ্যায়ত স্থাপন কৰা হ'ব এহেজাৰ খেলো ইণ্ডিয়া কেন্দ্ৰ Translation: One thousand Khelo India Kendra will be set up at the district level. Expected Caption: কেন্দ্ৰীয় ক্ৰীড়ামন্ত্ৰী কিৰেন ৰিজিজুৱে খেল আঁচনি হাতত লোৱাৰ কথা ঘোষণা কৰে Translation: Union Sports Minister Kiren Rijiju announces the launch of the sports scheme
 প্ৰতি দিন অগ্ৰণী খবৰ C	News Caption: দৰঙত পুনৰ ৩ জনৰ দেহত কোভিড-১৯ঃ সম্প্ৰতি চিকিৎসাধীন ৮৪ জন Translation: Covid-19 in 3 bodies again in Darang: 84 currently undergoing treatment Expected Caption: অগ্ৰণী খবৰ Translation: Breaking news (*photo feeds are not related to that current event*)

Fig. 3 Sample image-caption dataset

3.3.1 Encoder: Feature extraction

We use a convolutional neural network (CNN) to extract the image feature as an encoder. A convolutional neural network consists of three layers: convolutional layer, pooling layer and fully connected layer [26]. The basic functionality of the three layers is summarized into three distinct areas. The convolutional layer is connected to local regions of the input to determine the output. To perform down-sampling of the input and to reduce the number of parameters, the pooling layer is used. The convolutional and pooling layer is typically used for feature extraction. Finally, the fully connected layer performs the classification. The basic architecture of the convolutional neural network is shown in Fig. 4. We use the VGG-16 model for image feature extraction. It is a convolutional neural network for classification and detection.

3.3.2 Decoder: Textual representation

A recurrent neural network (RNN) is an excellent promise in natural language processing and speech processing tasks. RNN is a feed-forward neural network with internal memory where the previous state's output is fed into the current state. In image captioning, RNN is used as a decoder for generating textual description. The basic architecture of the recurrent neural network is shown in Fig. 5.

RNN fails if the information length goes beyond a measure. This problem is known as the vanishing gradient [13]. A forgetting mechanism has been proposed to resolve this problem named long short term memory (LSTM) [14]. LSTM is a special kind of recurrent neural network capable of learning long-term dependencies to address the vanishing gradient problem. It has different applications such as speech-to-text transcription, machine translation and other real-life applications [31]. It also plays a vital role in image captioning by solving the sequence problem. The architecture of the LSTM cell is shown in Fig. 6 and the definition of the LSTM cells is given below.

There are three gates in the LSTM cell to protect and control the cell state. First, it is decided what information is to be discarded from the cell state. This is decided by sigmoid layer called the "forget gate layer." Based on the input values of h_{t-1} and x_t, a number is generated between 0 and 1 in the cell state C_{t-1}, where 1 denotes "completely keep this" while a 0 denotes "completely get rid of this." The forget gate definition is shown in (1).

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \tag{1}$$

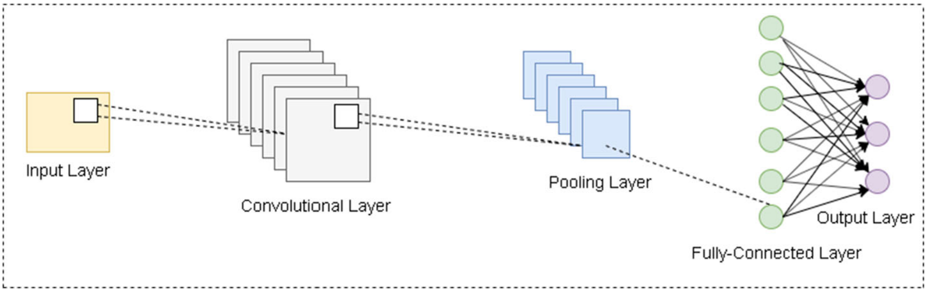


Fig. 4 Convolutional neural network

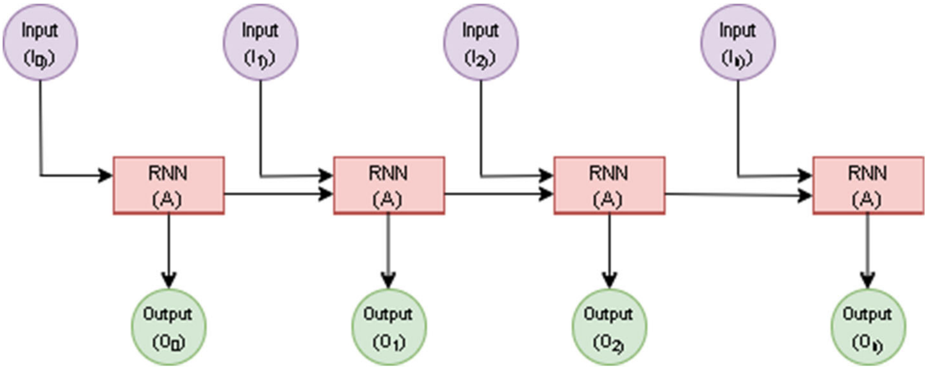


Fig. 5 Recurrent neural network

In the second step, it is decided what is to be stored in the cell state. First, a sigmoid layer called the "input gate layer" and decides which values will be updated. The input gate is defined in (2).

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{2}$$

Finally, the model needs to decide what is going to output gate. It controls the information of the current cell state and makes things visible. The definition of the output gate is shown in (3).

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \tag{3}$$

where,

x_t: input vector
h_{t-1}: hidden state at time stamp $t - 1$
h_t: hidden state at time stamp t
C_t: cell state at time stamp t
tanh: activation function

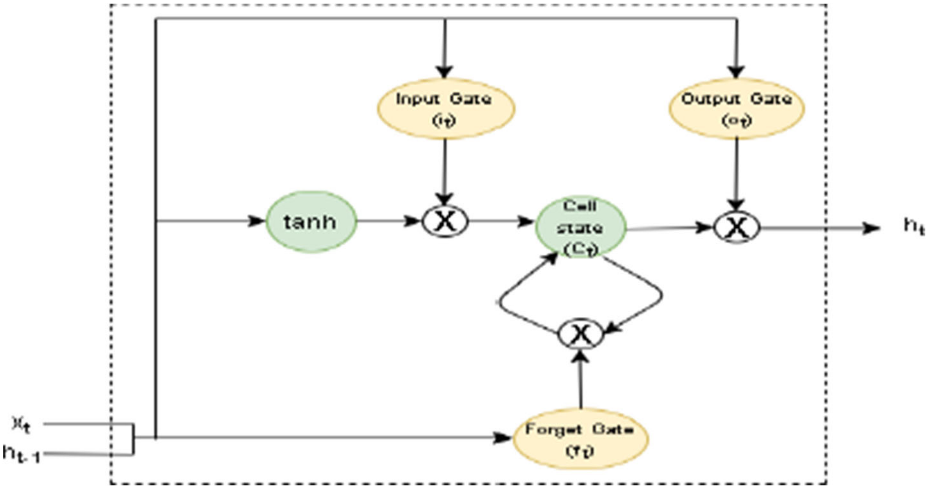


Fig. 6 Long short term memory cell

σ: activation function
f_t: forget gate
W_f: weight of forget gate
b_f: bias of forget gate
i_t: input gate
W_i: weight of input gate
b_i: bias of input gate
o_t: output gate
W_o: weight of output gate
b_o: bias of output gate
X: point-wise operation

3.4 CNN-LSTM architecture

The image caption generation task can be divided into two parts such as image feature extraction and generating an image description. The image and sequence of tokens are the input into the training model. The average length of the token is 10. Images are fed into CNN and the tokens are accepted by an embedding layer to generate the word embedding. We pre-trained our model using the VGG-16 [32] neural network to interpret the image contents.

Finally, the model uses one word at a time strategy to predict the image description. The image feature vector is fed into the LSTM layers. The output W$_{t-1}$ of the LSTM is provided as the input to the next LSTM layer which gives the output W$_t$ and the hidden state (h$_t$) at time t is a function of the hidden state (h$_{t-1}$) at time t−1 and the input at time t. We considered 11000 images for training the model and reserved 1000 images for development and 1000 images for testing. A large dataset is required to train the deep learning model. Since our dataset is admittedly small to train the caption generation system, we use a larger portion of the dataset for training in order to make the model reasonably robust. The developed model is heavily inspired by the architecture of “Show and Tell: A Neural Image Caption Generator”. The dataset mentioned in Table 1 is used in our task. Unrolling of the LSTM layer [37] to develop the image caption generation model is shown in the following (4), (5) and (6).

$$x_{-1} = CNN(I) \tag{4}$$

$$x_t \ = T_e W_n \quad t \in \{0, 1, ...N-1\} \tag{5}$$

$$P_{t+1} = LSTM(x_t) \quad t \in \{0, 1, ...N-1\} \tag{6}$$

In the LSTM based model shown in Fig. 7, the output of the previous layer at time t-1 is the input of the next layer at time t. W_n is the one-hot vector representing each word and W_0 is a special start character and W_n is a special stop character. The image with CNN from (4) and word by word embedding T_e from (5) are mapped to the same space. At time t= -1, the input image I is fed only once to convey LSTM about the image contents. Beam search iteratively inspects the k best sentences up to time t as candidates to generate sentences of size t + 1 and keep only the resulting best k of them. Beam search is a search strategy to give an optimized search for reducing memory requirements. It uses breadth-first search to build the search tree. A detailed description of the CNN-LSTM caption generation system is depicted in Fig. 7.

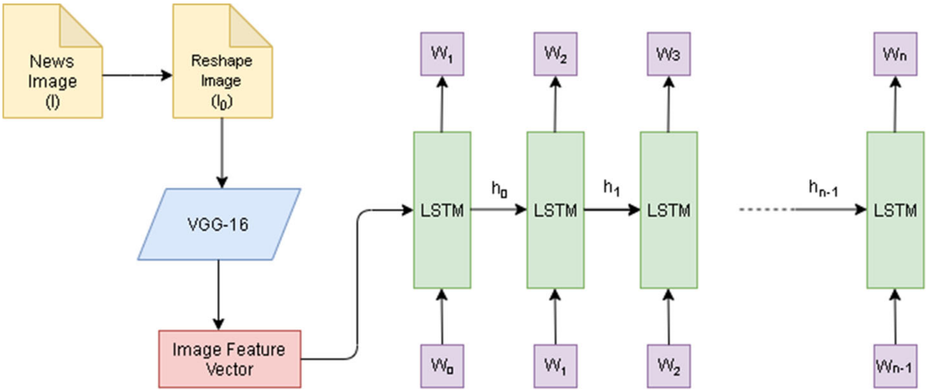


Fig. 7 News image caption generation model using CNN-LSTM

3.5 CNN-LSTM architecture with attention mechanism

In a sequence to sequence (seq2seq) model, the fixed-length context vector fails to remember long sentences. Therefore, to address this problem, an attention mechanism is used. In deep neural networks, the attention mechanism concentrates on a few relevant things and ignores the other parts. The model should know where to pay more attention and what would be negligible. The attention mechanism is helpful for the caption generation framework. Luong et al. [21] reported simplified attention mechanism proposed by [2]. The attention mechanism is broadly classified into two types. One is global attention and the other is local attention. When attention is paid to all source parts is known as global attention. Local attention is paid to only a few source positions. They referred to global attention as soft attention and local attention as a hard attention mechanism in their work. Local attention is less expensive and selects one part of the image to concentrate. This attention mechanism is also easier to train than global attention. An attention-based caption generation system is depicted in Fig. 8. The proposed CNN-LSTM with attention mechanism consists of the following components.

1. Encoder-CNN
2. Decoder-LSTM
3. Attention Mechanism-Local

For pre-processing the captions, we add "start" and "end" tags to every caption so that the model can understand the starting and end of the sentence. Since we use the VGG-16 model, the image is reshaped to 224*224 pixels. Our model is inspired by the architecture of "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". We employed a local attention mechanism to generate the image caption. We evaluated the captions using the beam search method. For selecting the final captions, we picked up top k predictions to feed into the model and again sort them using the probabilities returned by the model. The working principles of local attention mechanism [21] is described below. Figure 9 shows the description of the local attention mechanism.

Working principles of local attention mechanism:

1. Encoder produces hidden state of each element.
2. Calculate alignment scores between encoder states.

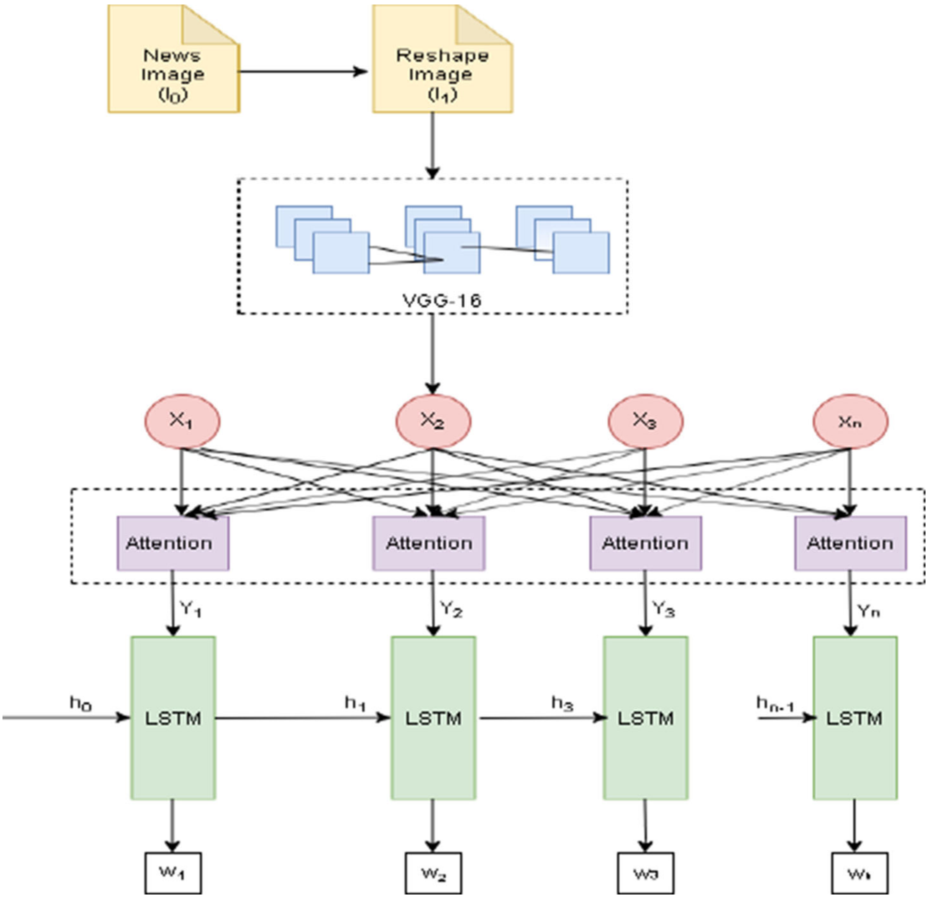


Fig. 8 News image caption generation model with attention mechanism

3. Softmax the alignment score for each hidden state.
4. Calculate the context vector by multiplying the encoder hidden states and the respective alignment scores.
5. The final step is decoding the output.

4 Experimental results and analysis

In this section, we cover the results and analysis of the system performance. Model evaluation can be quantitative or qualitative and often embrace both. For quantitative analysis, the BLEU and CIDEr metrics are used. Adequacy and fluency are two rating scales to assess our qualitative analysis.

4.1 Quantitative analysis

BLEU [27] and CIDEr [36] are two evaluation tools [5] used for evaluating the quality of generated captions. BLEU (stands for bilingual evaluation understudy) is mainly used in

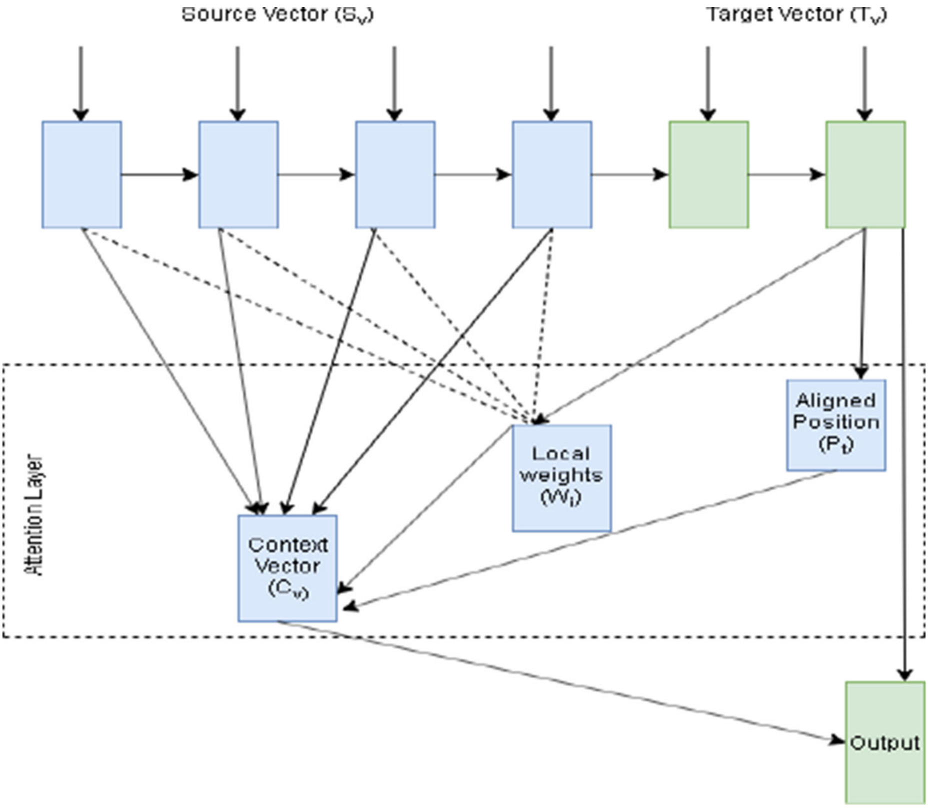


Fig. 9 Local attention mechanism

machine translations and CIDEr finds its use in image captioning. The evaluation range of BLEU is from 0 to 1. BLEU-1, BLEU-2, BLEU-3, and BLEU-4 are the individual N-gram scores where 1,2,3 and 4 represent 1-gram, 2-gram, 3-gram, and 4-gram. The formula used to derive the BLEU score is given below.

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

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

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

CIDEr is an evaluation metric explicitly developed for the image caption generation model. It measures the similarity between the automatically generated image description and the reference sentences manually annotated by annotators. CIDEr calculates consensus in image captions by performing a term-frequency inverse document frequency (TF-IDF) weighting for each n-gram. Where each sentence is considered as a document represents it in the form of a TF-IDF vector and the cosine similarity of the ground truth description to the model generated description as a score is calculated. We report the BLEU-1, BLEU-2, BLEU-3, BLEU-4 and CIDEr scores in Table 2.

We also plot the loss vs. epoch curve for both models. The loss curve is to calculate the loss value after training the model for the number of epochs. Generally, the loss value decreases when the number of epochs increases. We run both models for 40 epochs. As shown in Fig. 10 which is the loss curve of the CNN-LSTM model, the loss value starts with 4.2 and decreases up to 3.6. Again in the attention model shown in Fig. 11, the loss value decreases from 1.5 to 0.8. In both cases, the loss value decreases with the increasing number of epochs. From the above graph, we conclude that the attention-based learning rate is higher than the standard CNN-LSTM model.

4.2 Qualitative analysis

A human judge is a common way to assess the quality of automatic machine-generated texts. The automatically generated captions are judged in terms of adequacy and fluency rating. Adequacy and fluency are two rating scales to assess the qualitative analysis. How much meaning the target text can express as compared to the source text is called adequacy. The fluency is the measure of ability to express a grammatically correct target text. Qualitative analysis is a directional evaluation that gives the output direction without quantifying it. It includes interpreting the meaning, comparing, and contrasting the patterns. We have qualitatively evaluated our model performance based on 4 point scale rating. We engage two native speaker of Assamese as human annotators for qualitative analysis looking into fluency and adequacy. Tables 3 and 4 shows on a 4-point scale rating of adequacy and fluency respectively. Sample input and output for the image caption generation model are listed in Fig. 12. As can be seen from the sample input-output example, generated descriptions are pretty much close to the input images. However, still, there is a need for improvement in the machine-generated captions. The ground truth is that the quality of the caption for an image can be improved by training on a larger dataset with multiple annotated captions.

As shown in Fig. 12a, the model can show a quite impressive result. The machine-generated caption is meaningful and fluent. So, based on the point scale rating (Tables 3, and 4), both adequacy and fluency are 4. As shown in Fig. 12b, the model can also show an excellent result. In the second example, the machine-generated caption is meaningful and fluent. So, based on the point scale rating, both the adequacy and fluency are 4. As shown in Fig. 12c, the model can not identify the named entity as “Baghjan”. The human caption and machine caption is less meaningful and fluent. Based on the point scale rating, we can rate adequacy to 3, and fluency is 4. The system cannot identify the test image in Fig. 12d. The generated caption is not meaningful and flawless. Therefore, based on the 4 point rating scale, we rate adequacy to 1 and fluency to 1. We randomly choose 100 sentences from

Table 2 Evaluation metrics

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr
CNN-LSTM	30.5	23.2	19.4	9.4	44.4
CNN-LSTM with Attention Mechanism	38.4	27.6	20.8	11.1	49.5

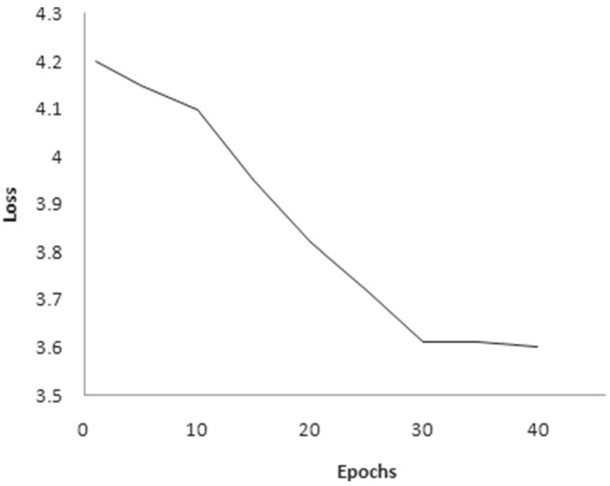


Fig. 10 Loss vs Epochs with CNN-LSTM mechanism

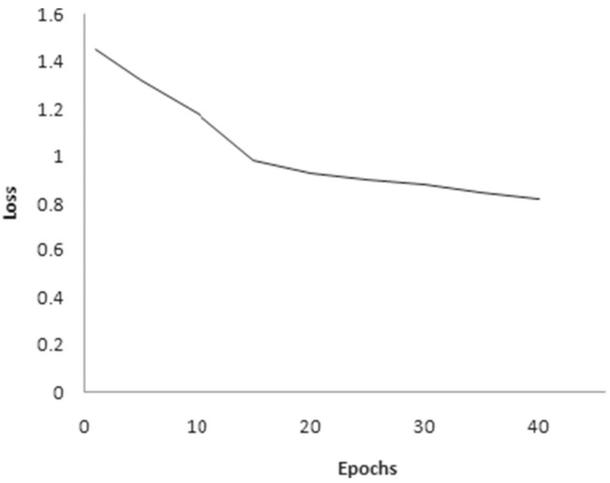


Fig. 11 Loss vs Epochs with Attention mechanism

Table 3 Adequacy on 4-point scale rating

Adequacy	Point scale
All meaning	4
Most meaning	3
Little meaning	2
No meaning	1

Table 4 Fluency on 4-point scale rating

Fluency	Point Scale
Flawless	4
Good	3
Disfluent	2
Incomprehensible	1

the test dataset to calculate the adequacy and fluency. From the machine-generated caption, most of the captions are flawless. Table 5 shows the adequacy and fluency evaluation scores of both CNN-LSTM and CNN-LSTM with an attention mechanism. From the human evaluation score, attention-based mechanism shows a better performance.

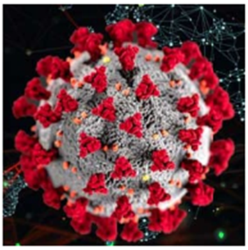
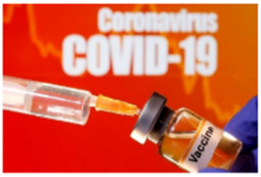
Input	Output
 A	Human Caption: ভাৰতত কোৰোনাত আক্ৰান্ত হৈ মৃত্যু হোৱা লোকৰ সংখ্যা বৃদ্ধি Translation: Corona death toll rises in India Machine Caption: ভাৰতত কোৰোনাত আক্ৰান্ত হৈ মৃত্যু হোৱা লোকৰ সংখ্যা বৃদ্ধি Translation: Corona death toll rises in India
 B	Human Caption: ৰাজ্যৰ স্বাস্থ্য মন্ত্ৰী ড০ হিমন্ত বিশ্ব শৰ্মাৰ সংবাদমেল Translation: State Health Minister Dr. Himanta Bishwa Sharma in a conference. Machine Caption: ৰাজ্যৰ স্বাস্থ্য মন্ত্ৰী ড০ হিমন্ত বিশ্ব শৰ্মাৰ সংবাদমেল Translation: State Health Minister Dr. Himanta Bishwa Sharma in a conference.
 C	Human Caption: বাঘজানত ভয়ংকৰ অগ্নিকাণ্ড Translation: Terrible fire in Baghjan Machine Caption: চন্দ্ৰযান ২ৰ ভয়ংকৰ অগ্নিকাণ্ড Translation: Terrible fire in Chandrayaan 2
 D	Human Caption: ক'ৰোনা মহামাৰীয়ে বিশ্বজুৰি সংকট সৃষ্টি কৰিছে Translation: Corona epidemic has created a global crisis Machine Caption: মৈৰাবাৰীত এযোৰ হাতমোজা Translation: A pair of gloves in Mairabari

Fig. 12 Sample 1 input and output

Table 5 Human evaluation results

Model	Adequacy	Fluency
CNN-LSTM	1.98	2.86
CNN-LSTM with Attention Mechanism	2.91	3.85

4.3 Discussion

There is a difference between traditional image captions and news image captions. News caption generation input consists of a news image that is accompanying the article. The news image always conveys more detailed information regarding the particular news. The conventional image captioning model describes the image properties to identify the object and enumerate the relationship with each other. As opposed to that, the output of the news caption generation model carries a piece of more detailed information that not only describes the semantic relationship but also expresses a meaningful summary regarding the particular news article. The main difficulty of news image caption generation is caption annotation. It is difficult to annotate an appropriate news image description corresponding to the article. Because many times, photo feeds are not related to that current event. Therefore, the caption is not evident from the image itself. The main difficulty of our work is the poor quality of captions present in the raw Assamese news articles. Sometimes, there is only one image with several news articles that is not suitable for the news caption generation model. One more unconventional phenomenon is that the image caption and news image are different from each other. Further, there are instances of some articles with only the logo images or file images but not relevant to the news event. As illustrated in Fig. 3, some of the annotated news captions are not relevant to the news image. So, as part of the pre-processing, the human annotators have to annotate each image with an appropriate caption that will be more relevant to the news event.

5 Conclusion and future research direction

This paper introduces an Assamese news image caption generation model using a deep neural network architecture using news images with one description per image collected from different local e-newspapers. Our first model is the CNN-LSTM based model for generating the image caption. We extracted image features using CNN and feed them into LSTM for further description generation. The model is pre-trained using the VGG-16 on the image-caption pair dataset. LSTM is combined with CNN to be trained with images and generates one word at a time. In the other model, we add one attention layer on top of the LSTM layer. Attention mechanism decides where to put attention and what information to analyze and plays an essential role in the next movement. We carried out both qualitative and quantitative analyses to evaluate model performances. For the quantitative analysis, the BLEU and CIDEr evaluation metrics are used. Further, it is shown that the attention based model works better in terms of the BLEU and CIDEr scores. A qualitative analysis of fluency and adequacy is also carried out. It is also observed that the attention-mechanism based model outperforms the CNN-LSTM based model. There are several problems to be addressed in the future research work. One of them is to attempt to include more varied and larger collection of images of various events into the dataset with multiple appropriate captions to generate a standard human-like caption. A full-fledged architecture might give us a delightful experience of the system performance in our real-life applications.

References

1. Amritkar C, Jabade V (2018) Image caption generation using deep learning technique. In: 2018 Fourth international conference on computing communication control and automation (ICCUBEA), IEEE, pp 1–4
2. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In Proc international conference on learning representations arXiv:1409.0473
3. Bai S, An S (2018) A survey on automatic image caption generation. Neurocomputing 311:291–304. Elsevier
4. Batra V, He Y, Vogiatzis G (2018) Neural caption generation for news images. In: Proceedings of the Eleventh international conference on language resources and evaluation (LREC 2018)
5. Chen X, Fang H, Lin T-Y, Vedantam R, Gupta S, Dollár P, Lawrence ZC (2015) Microsoft coco captions: Data collection and evaluation server. arXiv:1504.00325
6. Chen X, Lawrence Zitnick C (2015) Mind's eye: A recurrent visual representation for image caption generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2422–2431
7. Dhir R, Mishra SK, Saha S, Bhattacharyya P (2019) A deep attention based framework for image caption generation in hindi language. Computación y Sistemas 23:3
8. Fang H, Gupta S, Iandola F, Srivastava RK, Deng L, Dollár P, Gao J, He X, Mitchell M, Platt JC et al (2015) From captions to visual concepts and back. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1473–1482
9. Feng Y, Lapata M (2010) How many words is a picture worth? automatic caption generation for news images. In: Proceedings of the 48th annual meeting of the Association for Computational Linguistics, pp 1239–1249
10. Feng Y, Lapata M (2012) Automatic caption generation for news images. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(4):797–812. IEEE
11. Gorokhovatskyi O, Peredrii O (2018) Shallow convolutional neural networks for pattern recognition problems. In: 2018 IEEE Second international conference on data stream mining & processing (DSMP), IEEE, pp 459–463
12. Haripriya B, Srushti GM, Haseeb S, Prakash MM Image Captioning using Deep Learning
13. Hochreiter S (1998) The vanishing gradient problem during learning recurrent neural nets and problem solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 6(02):107–116
14. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
15. Holzinger A, Saranti A, Mueller H (2021) KANDINSKY Patterns–An experimental exploration environment for Pattern Analysis and Machine Intelligence. arXiv:2103.00519
16. Kamal AH, Jishan Md, Mansoor N et al (2020) TextMage: The Automated Bangla Caption Generator Based On Deep Learning. arXiv:2010.08066
17. Karpathy A, Fei-Fei L (2015) Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3128–3137
18. Kohakade AK, Emmanuel M (2014) Content based caption generation for images embedded in news articles. Int J Comput Appl 100(11):7–15
19. Lu X, Wang B, Zheng X, Li X (2017) Exploring models and data for remote sensing image caption generation 56(4):2183–2195. IEEE
20. Lu D, Whitehead S, Huang L, Ji H, Chang S-F (2018) Entity-aware image caption generation. arXiv:1804.07889
21. Luong M-T, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. arXiv:1508.04025
22. Mansimov E, Parisotto E, Ba JL, Salakhutdinov R (2015) Generating images from captions with attention. arXiv:1511.02793
23. Meetei LS, Singh TD, Bandyopadhyay S (2019) Extraction and identification of manipuri and mizo texts from scene and document images. In: Deka B, Maji P, Mitra S, Bhattacharyya DK, Bora PK, Pal SK (eds) PReMI 2019. LNCS, vol 11941. Springer, Cham, pp 405–414. https://doi.org/10.1007/978-3-030-34869-4_44
24. Meetei LS, Singh TD, Bandyopadhyay S (2019) WAT2019: English-Hindi translation on Hindi visual genome dataset. In: Proceedings of the 6th workshop on asian translation, pp 181–188
25. Miyazaki T, Shimizu N (2016) Cross-lingual image caption generation. In: Proceedings of the 54th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 1780–1790
26. O'Shea K, Nash R (2015) An introduction to convolutional neural networks. arXiv:1511.08458

27. Papineni K, Roukos S, Ward T, Zhu W-J (2002) BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp 311–318
28. Peng H, Li N (2016) Generating chinese captions for flickr30k images
29. Prajapati K, Wadekar S, Bobhate B, Mhatre A Auto-Caption Generation for News Images
30. Rahman M, Mohammed N, Mansoor N, Momen S (2019) Chittron: An automatic bangla image captioning system. Procedia Comput Sci 154:636–642. Elsevier
31. Sherstinsky A (2020) Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. Physica D: Nonlinear Phenomena 404:132306. Elsevier
32. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
33. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
34. Singh A, Meetei LS, Singh TD, Bandyopadhyay S (2021) Generation and evaluation of hindi image captions of visual genome. In: Maji AK, Saha G, Das S, Basu S, Tavares JMRS (eds) Proceedings of the international conference on computing and communication systems. Lecture Notes in Networks and Systems, vol 170. Springer, Singapore. https://doi.org/10.1007/978-981-33-4084-8_7
35. Soh M (2016) Learning CNN-LSTM architectures for image caption generation. Dept Comput Sci, Stanford Univ., Stanford, CA, USA, Tech. Rep
36. Vedantam R, Lawrence Zitnick C, Parikh D (2015) Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4566–4575
37. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: A neural image caption generator, pp 3156–3164
38. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning, pp 2048–2057
39. You Q, Jin H, Wang Z, Fang C, Luo J (2016) Image captioning with semantic attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4651–4659

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.