

# Image Caption Generation for Low-Resource Assamese Language

Prachurya Nath<sup>1</sup>, Prottay Kumar Adhikary<sup>1</sup>, Pankaj Dadure<sup>2</sup>,  
Partha Pakray<sup>1</sup>, Riyanka Manna<sup>3</sup>, Sivaji Bandyopadhyay<sup>1</sup>

<sup>1</sup>National Institute of Technology, Silchar, India

<sup>2</sup>University of Petroleum & Energy Studies, Dehradun, India

<sup>3</sup>ADAMAS University, Kolkata, India

{prachuryanath00, prottay71@gmail, krdadure, parthapakray,  
riyankamanna16, sivaji.ju.cse}@gmail.com

## Abstract

Image captioning is a prominent Artificial Intelligence (AI) research area that deals with visual recognition and a linguistic description of the image. It is an interdisciplinary field concerning how computers can see and understand digital images & videos, and describe them in a language known to humans. Constructing a meaningful sentence needs both structural and semantic information of the language. This paper highlights the contribution of image caption generation for the Assamese language. The unavailability of an image caption generation system for the Assamese language is an open problem for AI-NLP researchers, and it's just an early stage of the research. To achieve our defined objective, we have used the encoder-decoder framework, which combines the Convolutional Neural Networks and the Recurrent Neural Networks. The experiment has been tested on Flickr30k and Coco Captions dataset, which have been originally present in the English language. We have translated these datasets into Assamese language using the state-of-the-art Machine Translation (MT) system for our designed work.

**Keywords:** Caption Generation, Low-resource Language, Attention, Assamese.

## 1 Introduction

Over 24 million native speakers speak the Assamese language in the north-eastern part of India. It is an eastern Indo-Aryan (Indic) language which is the official language of India's Assam state. Assamese is an indigenous Indo-Aryan language which has been influenced in vocabulary, phonetics, and structure by the region's close association with Tibeto-Burman dialects. Its grammar is notable for its highly inflected forms, different pronouns,

plural noun markers, and honorific and non-honorific constructions. The Assamese script is very close to the Bengali script. Assamese, like English, is written from left to right.

The Assamese literary tradition can be traced back to the 13th century. In the 16th century, prose texts, most notably buranjis (historical works), began to appear. The Assamese alphabet (Assamese, Oxomiya bornomala) is shown in Fig. 1 which is the Bengali-Assamese script used in the Assamese language. Other north-eastern languages that are using the script include Bodo (now Devanagari), Khasi (now Roman), Mising (now Roman), Jaintia (now Roman), and others. The Kamarupi script was used to create it. Since Fifth century Umachal/Nagajari-Khanikargaon rock inscriptions written in an eastern variant of the Gupta script, the script has evolved continuously, with significant influences from the Siddha script in the 7th century (Saharia and Konwar, 2012). The current format is identical to the Bengali alphabet with the exception of two letters, (ro) and (vo); and the letter (khya) has progressed into an independent consonant with its own phonetic quality, however in the Bengali alphabet it is a conjunct of two letters.

Attempting to make computers mimic humans' ability to interpret the visual world is one of the long goal of artificial intelligence researchers. Even though significant advancement have been made in numerous computer vision tasks, for example, attribute classification (Lampert et al., 2009), object identification (Felzenszwalb et al., 2009), action classification (Maji et al., 2011), scene recognition (Zhou et al., 2014), and image classification (Krizhevsky et al., 2012), the field of Natural Language Processing have seen recent huge advances with the addition of transformer ar-

VOWELS (SWARVARNA)												
অ	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও	ঔ		
o	a	i	ī	u	ū	ri	e	oi	ū	ou		
CONSONENETS (BYANJANVARNA)												
ক	খ	গ	ঘ	ঙ	চ	ছ	জ	ঝ	ঞ	ট	ঠ	
k	kh	g	gh	ng	s	sh	z	zh	y	t	th	
ড	ঢ	ণ	ত	থ	দ	ধ	ন	প	ফ	ব	ভ	
d	dh	n	t	th	d	dh	n	o	ph	b	bh	
ম	য	ৰ	ল	ৱ	শ	ষ	স	হ	ক্ষ	ড়	ঢ়	য়
m	z	r	l	w	śa	ṣa	sa	h	khy	r	rh	y
NUMBERS (SANKHYA)												
০	১	২	৩	৪	৫	৬	৭	৮	৯			
quinno	ek	dui	tini	sari	pas	soy	xat	ath	no'			

Figure 1: Assamese alphabets

chitectures. Moreover, allowing a computer to automatically describe an image in human language is a comparatively new task.

Image caption generation is the process of describing the visual information of an image based on the objects and actions depicted in the image using a machine’s visual perception and a language model. The study of how computers can apprehend digital images and videos as well as describe them in a language that humans can understand is an interdisciplinary field.

Recent advancements in the field of Natural Language Generation (NLG) have helped the advancements of a plethora of fields like Machine Translation, Text Summarization, Answer Generation, and Image Captioning (Min et al., 2021). The inclusion of several pre-trained transformer-based language models such as BERT (Devlin et al., 2019) have taken the Natural Language Processing (NLP) to new heights. The interpretation of an image is highly dependent on acquired image features. In the prior studies, there are two approaches that have been taken into consideration to accomplish this task (Wang et al., 2020): one that uses a statistical probability language model to generate handcrafted features, and another one uses the neural net-

work models based on encoder-decoder language model to extract the deep features.

In this paper, we propose an encoder-decoder framework for creating image captions in Assamese. The proposed model is based on a separate language model and a visual understanding machine. The rest of this paper has placed out as follows: Section 2 presents the works that has common factors with our work. The data used in our experiment has been discussed in 3 section. 4 section contains the procedures we used to prepare our system. The results, advantages and drawbacks have been discussed in 5 section. Finally, in Section 6, the paper is concluded with a discussion on future work.

## 2 Related Works

Most of the image caption generation systems comprised of rudimentary vision-based signifiers and language models which have been used during the early stage of the research. These systems mainly includes rule-based and hand-coded approaches. These systems only worked on a limited set of images. In recent time, the image captioning systems produced significantly improved results, following the same deep learning-based architecture as machine translation, as deep learning meth-

ods enhanced. These works were using the same encoder-decoder framework and framed image captioning as a text-to-image translation. CNN was used to encode images and RNN was used to decode the images into sentences in these systems.

Vinyals et al. (2015) (Vinyals et al., 2015) used CNN as an encoder to encode images and RNN-LSTM to decode image features into text, where image captioning is defined as predicting the probability of a sentence based on the input image feature. The most simple LSTM-based captioning architecture is based on a single-layer LSTM. During training, input words are taken from the ground-truth sentence, while during inference, input words are those generated at the previous step. Donahue et al. (2014) (Donahue et al., 2017) provide both image and text features to the sequential language model at each time step, rather than inputting image features to the system at the start. The encoder-decoder framework's next advanced version is an attention guided framework. Xu et al. (Xu et al., 2015) proposed the first attention mechanism in image caption generator. The encoder-decoder framework is more focused on the salient region of an image while generating an image description. It is a method that allows you to weight different areas of an image differently. It can, for example, add more weights to an image's important region. The attention model developed by them involved assigning weights to a random portion of an image. As a result, some critical aspect of an image was overlooked in order to generate a caption. To address this limitation, You et al. (You et al., 2016) developed a semantic attention model that focuses on linguistically significant objects or action in the image. In the preceding attention mechanism, the model forces visual attention to be active for every word, even those that do not explain visual information. Stop words such as 'the', 'of', and so on do not explain the image object. To address this issue, Lu et al. (Lu et al., 2018) developed an adaptive attention mechanism, which automatically determines whether to rely on the visual signal or the language model. Whenever the adaptive attention model starts paying attention to a visual signal, it will automatically decide

which part of the image to focus on.

Vaswani et al (Vaswani et al., 2017) described the fully-attentive paradigm which has changed the way people think about language generation completely. The Transformer model was fully embraced as the de-facto architecture for several language understanding tasks, as well as the groundwork for other advancements in NLP, such as BERT and GPT.

The Transformer architecture has been used for image captioning because it can be viewed as a sequence-to-sequence problem. A masked self-attention operation is applied to words in the standard Transformer decoder, followed by a cross-attention operation. Words serve as queries, and the final encoder layer's outputs serve as key / value, as well as an ultimate feed-forward network. Improvement of language generation and visual feature encoding have also been proposed.

The North-East of India is one of the country's most linguistically and culturally diverse regions. Every state has its own culture, language, and customs. Languages serve as a link between people and aid in the formation of bonds. The languages are mostly divided into three groups: Indo-Aryan, Sino-Tibetan, and Austro-Asiatic. Assamese, Bengali, English, Hindi, Manipuri, and Nepali are the most widely spoken languages in the North-east. As a result, the Northeast is also known as India's multilingual and multicultural region. For natural languages, a large number of different NLP applications is being developed in India, as well as across the world. As Saiful Islam et al. (Devi and Purkayastha, 2018) described, there are only a few NLP applications for NE languages have been developed in India.

Natural Language Processing in Assamese is being worked on in a number of different ways. Assamese is a computationally underdeveloped language, and NLP study is still in its early stages. Works have mainly been carried out in the fields of Machine Translation as we can see in the work of English to Assamese using Statistical Machine Translation(SMT) (Singh et al., 2014). Laskar et al worked on multi-modal translation using both textual and visual features (Laskar et al.,

2019). Other works have been done in the field of Automatic Speech recognition by Agarwalla et al (Agarwalla and Sarma, 2016) and Supervised named entity recognition in Assamese language (Talukdar et al., 2014). In this piece of writing, we are suggesting an encoder-decoder framework for writing Assamese image captions. The suggested model is based on a unique language model and a machine that can understand visuals.

### 3 Dataset

A number of datasets are available for high-resource languages such as English, Hindi, etc., to carry out the experiments in image caption generation. For low-resource languages, the unavailability of sufficient data is the prime challenge faced by the researcher. In the low-resource Assamese language, there is no data available for the task of image caption generation. In this proposed work, we have generated the data using the Translator Cognitive Service provided by Microsoft Azure. Herein, we have used the Flickr30k and MS COCO Dataset (Lin et al., 2014) which are initially available in English. Moreover, we have used the machine translation systems (Translator Cognitive Service provided by Microsoft Azure) to translate the captions of these datasets into the Assamese language. The generated dataset is pictorially depicted in Fig. 2. The designed datasets differ in several ways, including the number of images, the number of captions per image, the format of the captions, and the size of the images.

#### 3.1 Flickr30k

Flickr30K (Young et al., 2014) is one of the most popular datasets used for automatic image description and grounded language understanding. It includes 30000 Flickr images and 158000 human-annotated captions. Initially, it does not provide predefined image splits for training, testing, and validation. Herein, the researchers are free to select their own numbers for training, testing, and validation splits as per the requirements.

#### 3.2 MS COCO Dataset

The Microsoft COCO (MS COCO) Dataset (Chen et al., 2015) is a popular massive

dataset used for several tasks like image recognition, object segmentation, and captioning dataset. It contains multiple objects per class, with over 300,000 images, over 2 million instances, 80 object categories. In this dataset, 5 captions have been available for each input image.

Table 1: Dataset Description

Name of Dataset	Train	Test
Flickr30k	29783	2000
Coco17	118287	5003
Combined	148070	7003

In this work, we have used Flickr30k and MS Coco datasets. For more promising results, we have combined both datasets and analyzed the performance in terms of domain-independent perspective. Flickr30k consisting of 31,783 images and Microsoft Coco(MS-COCO) 2017 Captions dataset which has 118287 training images. Both Flickr30k/Coco Captions come with 5 human-annotated captions for each image. Table 1 shows the statistics of the used dataset.

### 4 Framework

The architecture of the proposed model is shown in Fig. 3 which primarily relies on the encoder-decoder mechanism. In the proposed model, image features are encoded using a convolutional neural network, and the image captions (word sequences) are encoded using a recurrent neural network. Later, the encoded image is passed to a text feature decoder which predicts the caption word by word. As it generated each word of the caption, the model used attention to focus on the most important part of the image.

#### 4.1 Image Features Encoder

We use transfer learning to preprocess the raw files, using a CNN-based system which has already been trained. The images are fed into this process, which produces encoded image vectors that capture the image’s essential features. For image feature extraction, we have used pre-trained VGG16 (Simonyan and Zisserman, 2015) and EfficientNetB3 (Tan and Le, 2020). It was trained using the ImageNet dataset. Historically, neural networks with



Figure 2: Overview of the source data

many layers have performed well in pattern recognition. Apart from this, these models are also suffer from the overfitting and difficult to optimise. Residual CNNs are comprised of many layers with interconnections between them. Identity mapping is decided to carry out by these connections. VGG16 has 16 layers and is simpler than EfficientNetB3. EfficientNetB3 are simple to optimise, and their performance improves as network depth increases. We used only the encoded image features produced by the hidden layers and discarded the pretrained models' final output layer because it contains the final output of classification.

#### 4.2 Word Sequences Encoder

We tokenize our sentences with Tensorflow and extract the tokens from the top 25000 words. The tokens are then passed through an Embedding layer with embedding size=256 and an RNN based on Gated Recurrent Units (GRU). Kyunghyun Cho et al. (Cho et al., 2014) introduced GRU, which has been successfully used for machine translation and sequence generation.

The Generalized Recurrent Neural Network (GRU) is an improved version of the Recurrent Neural Network. The update gate and reset gate are used in standard RNN to solve the

vanishing gradient problem. These two gates are in charge of the cell's behaviour. The memory cell at the heart of the GRU model stores information about each time step (what input has been observed up to this point). The update and reset gates are the vectors which determines the forwarding of the specific information to the output. The two gates have been developed to save input from earlier time steps without losing it and to eliminate data which is unrelated to the forecast.

The main distinction between Gated Recurrent Units (GRU) and Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) is that GRU's bag has two gates: reset and update, whereas LSTM's bag has three gates: input, output, and forget. Because GRU has fewer gates than LSTM, it is less complex. GRU is 29.29% faster than LSTM for same dataset in terms of model training speed; and in terms of results, GRU will out-compete LSTM in the case of long text and comparatively tiny data sources, but will fall well short in other instances.

#### 4.3 Attention Mechanism

For our experiments, we have used Bahndau Attention, as described in the research articles 'Neural Machine Translation by Jointly

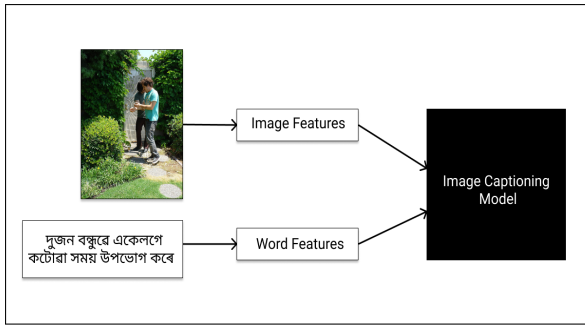


Figure 3: Model Overview

Learning to Align and Translate’ (Bahdanau et al., 2016). The attention features shape for VGG16 is 49, while it is 100 for Efficient-NetB3. The context vector generated by the pretrained model’s last hidden layer is passed to attention layer. GRU obtains the context vector as an input and produces an image description. This architecture outperforms traditional CNN and RNN architectures that use Long Short Term Memory (LSTM) as a decoder.

At each keyframe, the Attention module receives the encoded image along with the previous timestep’s hidden state from the Decoder. It generates an Attention Score, which gives each pixel in the encoded image a weight. The higher a pixel’s weight, the more likely it is that the word will be output at the next timestep. For example, if the target output sequence is A boy is kicking the ball, the boy’s pixels in the photo are highlighted when generating the word boy, while the ball’s pixels are highlighted for the word ‘ball’.

Attention is the process of focusing on a distinct aspect of information whilst dismissing other apparent information. It’s a way of telling the model where to focus in order to generate the corresponding word instead of the entire image. The decoder pays specific attention to some regions of the image at time  $t$ , on the basis of the hidden state, and by the use of spatial image features, it measures context vector.

#### 4.4 Caption Generation

The caption generator is composed of a simple decoder with a Dense layer and Rectified Linear Unit (ReLU) activation. The dense layer, which also includes attention weights, receives the output of the picture feature en-

coder. The dense layer generates a softmax prediction of the next word in the sequence for each word in the vocabulary, then chooses the word with the highest probability. Instead of raw photos, we pass these encoded image attributes into our Image Caption algorithm. The target captions for each encoded image are also passed in. By decoding visual information, the model attempts to predict captions that compliment the intended caption. During the training phase, we use the Teacher forcing method to predict the next word where the target word is passed as the next input to the decoder. This procedure is repeated until a final token is generated.

## 5 Results

The results evaluation of the proposed approach for Assamese language is quantitatively and qualitatively challenging task. The system generated results values are remarkable and it set benchmark for other existing systems for the task of caption generation in Assamese language. There are a variety of evaluation metrics used in image captioning tasks that can be found in the literature. The BLEU score (Papineni et al., 2002) is the most commonly used metric. In addition to this, the Rouge score (Lin, 2004) is also one of the popular metric to computes the performance of the image caption generation system. These metrics works by comparing a system generated captions with a set of reference summaries.

The test sets of Flickr30k and COCO 2017 datasets contain 2000 and 5000 test images, to evaluate the proposed model’s performance. For the test dataset, BLEU has been recorded. We also experimented with our combined dataset, which contains 150k images in the training set and 7000 test images in the test set. We keep track of both our BLEU and ROUGE and presented the scores in Table 2.

The majority of the images in this dataset, feature’s human subjects with captions that are nearly identical. As a result of being trained on a large number of similar human subjects, the model during testing is unable to distinguish and describe non-human subjects. Machine translation of English captions to Assamese language has some limitation to translate compound sentences. The combined sys-

Table 2: Evaluation scores

Name of Dataset	Model	BLEU	Rouge
Flickr30k	VGG16	0.2833	0.1011
	EfficientNet	0.3084	0.1137
Coco17	VGG16	0.2694	0.1054
	EfficientNet	0.2677	0.1049
Combined	VGG16	0.2134	0.0778
	EfficientNet	0.2389	0.0889

tem was supposed to give a better accuracy in terms of BLEU and rouge score. But, they surprisingly gave a bit worse result as both the datasets contained different type of images so combining both the dataset was not good for our system. Although, it leaves us a space for future to work better how to get better results after the combination of two datasets.

As shown in the Table 3, we get a fairly close match to the reference captions. It lacks in areas where the word does not appear frequently, so applying one shot learning for objects that doesn't contain many samples. In our results, the distinction between source and predicted captions is that some sources contain a detailed definition of the image. Moreover, the designed system tries to give the overall details of the image. Another limitation of the proposed approach is that the caption mainly focuses on one main area, as simplified version of Bahndau attention is use.

- The caption mentioned in Table 3 for the first image, the scissor is incorrectly referred to as sunglasses as the dataset contains a number of images of people with glasses.
- For second image, the proposed method analyses the different objects occurring in the image and attempts to predict the caption for the same. The generated caption is slightly confusing and semantically incorrect.
- Moreover, the source caption of the fourth image is overly detailed and the prediction is reasonable.
- In the third and fifth images, there is too much depth on the source caption.



Some drawbacks can be solved by the use of multi head attention which can help to solve this problem by focusing on more than one region. Also, transformers have a long way to go in the field of image captioning as the data is not always processed in the same order by transformers, and the attention mechanism provides context for any position in the input sequence.

## 6 Conclusions and Future Work

In this paper, we have proposed an encoder-decoder framework for creating image captions in Assamese. We looked at the Assamese alphabet and its presence in the Devanagari script, drawing parallels to the Bengali language. The current model is based on a separate language model and a visual understanding machine. Our primary focus was on translating English sentences, but in the long run, we are motivated to create a Gold Dataset for Assamese image captions. With the sudden rise in the use of Transformer-based frameworks in Machine Translation and other NLP tasks, image captioning using Attention-based Transformers could be a good experiment to investigate in the future.

Our main motivation was to create a benchmark model for Image Captioning in a low-resource language like Assamese for the first time. We hope that other machine learning researchers will work in this area to develop better models and improve the system's functionality and accuracy which may benefit many coming researchers. In the future, we'll explore the possibility of using a combination of different encoder and decoder architectures to enhance the result even further. We'll also experiment with different sampling techniques to see if we can eliminate the bias toward certain phrases. The gain of this experiment can be added to the research on Assamese image captioning and treated like a baseline model for further studies and future research.

Table 3: Caption Generated using Proposed Approach

Input Image	Captions
	<p><b>Source</b>                      As: এজন মানুহে কিছুমান কেঁচিৰ হেঙেলৰ মাজেৰে চাই আছে।                      En: A man is looking through the handle of a scissor.</p> <p><b>Predicted</b>                      As: চানপ্লাছ পিন্ধা মানুহজনে সক্ৰিয়ভাৱে এক গুৰুতৰ দেখাইছে।                      En: The man in sunglasses actively looks serious.</p>
	<p><b>Source</b>                      As: নিৰ্মাণ কৰ্মীসকলে বাহিৰত পাইপ সামগ্ৰী একত্ৰিত কৰে।                      En: Construction workers assemble pipe material outside.</p> <p><b>Predicted</b>                      As: নিৰ্মাণ কৰ্মীসকলে চহৰৰ মাজভাগত বেৰ জাঁপ প্ৰদৰ্শন কৰে                      En: Construction workers demonstrate wall jumps in the heart of the city</p>
	<p><b>Source</b>                      As: এজন মানুহে হ্ৰদৰ কাষৰ শিলৰ পথত এখন বাইক বৈ আছে।                      En: A man is waiting on a bike on a rock path near the lake.</p> <p><b>Predicted</b>                      As: বাইক চলাই থকা হেলমেট পিন্ধা এজন ব্যকত।                      En: A person wearing helmet riding bike</p>
	<p><b>Source</b>                      As: কিছুমান গছৰ সনুখত পোলকা ডট চাৰ্ট আৰু চশমা পিন্ধা এগৰাকী যুৱতী।                      En: A girl wearing a polka dot shirt and glasses in front of a tree.</p> <p><b>Predicted</b>                      As: ফটো তুলিবলৈ পোজ দিয়া এগৰাকী মহিলা                      En: A woman posing for a photo</p>
	<p><b>Source</b>                      As: ক'লা ৱেটচুট পিন্ধা এজন ব্যক্তিয়ে অকলে দৌত চাৰ্ফিং কৰে                      En: A person in a black wet suit is alone surfing in waves.</p> <p><b>Predicted</b>                      As: এজন মানুহে সাগৰৰ পাৰৰ সৈতে দৌ চলাই আছে।                      En: A man is riding a wave along the beach</p>



## References

- Swapna Agarwalla and Kandarpa Kumar Sarma. 2016. [Machine learning based sample extraction for automatic speech recognition using dialectal assamese speech](#). *Neural Networks*, 78:97–111. Special Issue on "Neural Network Learning in Big Data".
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. [Neural machine translation by jointly learning to align and translate](#).
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. [Learning phrase representations using rnn encoder-decoder for statistical machine translation](#).
- Maibam Devi and Bipul Purkayastha. 2018. [Advancements on nlp applications for manipuri language](#). *International Journal on Natural Language Computing*, 7:47–58.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. 2017. [Long-term recurrent convolutional networks for visual recognition and description](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):677–691.
- Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. 2009. [Object detection with discriminatively trained part-based models](#). *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. [Long short-term memory](#). *Neural computation*, 9:1735–80.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. [Imagenet classification with deep convolutional neural networks](#). *Advances in neural information processing systems*, 25.
- Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. [Learning to detect unseen object classes by between-class attribute transfer](#). In *2009 IEEE conference on computer vision and pattern recognition*, pages 951–958. IEEE.
- Sahinur Rahman Laskar, Rohit Pratap Singh, Partha Pakray, and Sivaji Bandyopadhyay. 2019. [English to Hindi multi-modal neural machine translation and Hindi image captioning](#). In *Proceedings of the 6th Workshop on Asian Translation*, pages 62–67, Hong Kong, China. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- 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 *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.
- Di Lu, Spencer Whitehead, Lifu Huang, Heng Ji, and Shih-Fu Chang. 2018. [Entity-aware image caption generation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4013–4023, Brussels, Belgium. Association for Computational Linguistics.

- Subhransu Maji, Lubomir Bourdev, and Jitendra Malik. 2011. Action recognition from a distributed representation of pose and appearance. In *CVPR 2011*, pages 3177–3184. IEEE.
- Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2021. [Recent advances in natural language processing via large pre-trained language models: A survey](#). *CoRR*, abs/2111.01243.
- 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*, ACL '02, page 311–318, USA. Association for Computational Linguistics.
- Navanath Saharia and Kishori M Konwar. 2012. [LuitPad: A fully Unicode compatible Assamese writing software](#). In *Proceedings of the Second Workshop on Advances in Text Input Methods*, pages 79–88, Mumbai, India. The COLING 2012 Organizing Committee.
- Karen Simonyan and Andrew Zisserman. 2015. [Very deep convolutional networks for large-scale image recognition](#).
- Moirangthem Tiken Singh, Rajdeep Borgohain, and Sourav Gohain. 2014. An english-assamese machine translation system. *International Journal of Computer Applications*, 93:1–6.
- Gitimoni Talukdar, Pranjal Protim Borah, and Arup Baruah. 2014. [Supervised named entity recognition in assamese language](#). In *2014 International Conference on Contemporary Computing and Informatics (IC3I)*, pages 187–191.
- Mingxing Tan and Quoc V. Le. 2020. [Efficientnet: Rethinking model scaling for convolutional neural networks](#).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#).
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. [Show and tell: A neural image caption generator](#).
- Haoran Wang, Yue Zhang, and Xiaosheng Yu. 2020. An overview of image caption generation methods. *Computational intelligence and neuroscience*, 2020.
- 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 *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France. PMLR.
- Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. 2016. [Image captioning with semantic attention](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4651–4659, Los Alamitos, CA, USA. IEEE Computer Society.
- 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.
- Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27.