An Automatic Bangla Image Captioning System



Department of Electrical & Computer Engineering

North South University

**Student Name (Student ID)**

1.Milon Hossain (1610497042)

2. Md. Musfiqur Rahman (1620277042)

3. Kh Sadik (1621536042)

4. Fahad Ahmed Pranto (1612607042)

30-09-2020

Cse499A – Senior Design Project

Mohammad Ashrafuzzaman Khan

# Abstract

Automatic image caption generation aims to produce an accurate description of an image in natural language automatically. It is a computationally challenging computer vision task which requires sufficient comprehension of both syntactic and semantic meaning of an image to generate a meaningful description. In recent years, with the rapid development of artificial intelligence, image caption has gradually attracted the attention of many researchers in the field of artificial intelligence and has become an interesting and arduous task. Image caption, automatically generating natural language descriptions according to the content observed in an image, is an important part of scene understanding, which combines the knowledge of computer vision and natural language processing.

We know Bangla is the 5th most widely spoken language in the world. While there are many established data sets exists to image annotation in English, no such resources exist for Bangla yet. For this reason, we are trying to develop an automatic image captioning system in Bangla. A deep neural network-based image captioning model was proposed to generate image description. The model employs Convolutional Neural Network (CNN) to classify the whole dataset, while Recurrent Neural Network

(RNN) and Long Short-Term Memory (LSTM) captured the sequential semantic representation of text-based sentences and generate pertinent description based on the modular complexities of an image.

Moreover, to address the data set availability issue, a collection of 16,000 Bangladeshi contextual images has been accumulated and manually annotated in Bangla. This data set is then used to train our model which integrates a pre-trained VGG16 image embedding model including LSTM layers. The model is train to predict a caption when we will put an image as an input to our model. The result will show that the model has successfully been able to learn a working language model and to generate captions of images in many cases.

# Table of Contents

Table of Contents

[Abstract 2](#_Toc52390732)

[Table of Contents 3](#_Toc52390733)

[1 Introduction 5](#_Toc52390734)

[2 Related Work 6](#_Toc52390735)

[2.1 Chittron 6](#_Toc52390736)

[2.2 Oboyob 6](#_Toc52390737)

[3 Background 7](#_Toc52390738)

[3.1 Overview of automated image captioning system 7](#_Toc52390739)

[3.1.1 Dataset 7](#_Toc52390740)

[3.1.2 Encoder 7](#_Toc52390741)

[3.1.3 Decoder 7](#_Toc52390742)

[3.1.4 Attention 8](#_Toc52390743)

[3.1.5 Beam Search 8](#_Toc52390744)

[3.2 Limitations and Challenges 8](#_Toc52390745)

[3.2.1 Limitations of other system 8](#_Toc52390746)

[3.2.2 Improvements in our system 9](#_Toc52390747)

[4 Methodology 9](#_Toc52390748)

[4.1 Convolutional Neural Network (CNN) 9](#_Toc52390749)

[4.2 Recurrent Neural Network (RNN) 9](#_Toc52390750)

[4.3 Gated Recurrent Unit (GRU) 10](#_Toc52390751)

[4.4 Long Short-Term Memory (LSTM) 10](#_Toc52390752)

[4.5 Image Captioning Evaluation Metrics 10](#_Toc52390753)

[4.6 Software Requirement 11](#_Toc52390754)

[4.7 Deep Learning Models 11](#_Toc52390755)

[5 Results and discussion 12](#_Toc52390756)

[6 Conclusions and recommendations 12](#_Toc52390757)

[7 Acknowledgements 13](#_Toc52390758)

[8 References 13](#_Toc52390759)

# 1 Introduction

The challenge of creating a natural language definition of an image applies to automated image captioning. This challenge is harder than the task of image detection and object recognition, because it not only requires objects within the image to be detected, but also requires their connection, language and behavior portrayed in the image to be detected. In addition, the expected data must be covered into a natural language that is comprehensible to humans. The biggest challenge is the problem of recognizing the graphical evidence that comes naturally to humans. An effective use for an automated image captioning application is to support visually disable individuals by providing them with data and the content of an image about automated image captioning in natural language. In search engines, another application is that photos can be searched by sentence fragments. Apart from the practical applications, image captioning requires the machine learning model to learn image understanding which is a significant computer vision challenge. It is possible to further apply the image captioning model to video captioning, which has many practical uses, including warning systems to improve protection.

While an image may be very well be worth “a thousand word”, it is hardly ever practical to describe an image with so many. Instead, what is important in many applications is an adequate description of an image comprising the essential information. The automated methods of image captioning are therefore targeted at doing exactly that, and have already had significant consequences in different fields, such as image search. Moreover, it has the potential to influence positive changes in many different areas, including software for disabled individuals, surveillance and security, human-computer interaction etc.

This project accomplishes this task using deep neural networks. By learning knowledge from image and caption pairs, the method can generate Bangla image captions that are usually semantically descriptive and grammatically correct.

# 2 Related Work

This section of the paper presents about two existing approaches for Bangla image captioning system. Here we will discuss related research efforts and explain why we focus on this topic. Basically, we are going to highlight two research paper’s idea on automated image captioning system which had been done using Bangla language and name of the papers are “Chittron” and another one is “Oboyob”.

## 2.1 Chittron

This paper was highlighted on An Automated Bangla image captioning system. A deep learning model was built in this project which was captioning an image automatically. The model uses the Convolution Neural Network (CNN) to identify the entire dataset, while the sequential semantic representation of text-related sentences was captured by Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) and the related description was created based on the modular complexities of a picture. A collection of 16,000 Bangladeshi contextual image has been accumulated and manually annotated in Bangla and this data set is then use into a model. The model is trained to predict the caption and result show that the model has successfully been able to learn a working language and generate a caption accurately.

## 2.2 Oboyob

The main idea of this paper is A descriptive definition of an image is necessary for summarizing the content of the image in a way that explain the story without digging into unnecessary details. Various Deep learning system like CNN, LSTM, VGG16 model had been followed in this paper. For evaluate and caption quality four more approaches also applied. They are “BLEU”, “METEOR”, “ROGUE” and “CIDER”. These terms are used to compute the length between reference caption and hypothesis caption. Here main problem was Bangla language does not prefer pre-trained word embedding. Finally, this model gives an accurate output or captioning the image appropriately.

# 3 Background

In this section, we give the overview of our automated image captioning system and closely monitor the relevant works limitations, challenges. We will also focus on the improvements in our automated image captioning system.

## 3.1 Overview of automated image captioning system

The system that we want to develop is based on the encoder and decoder architecture. We also used attention on attention framework, beam search.

### 3.1.1 Dataset

The dataset used for training and testing this model is BanglaLekha-ImageCaptions. This dataset has about 9115 images. The images are human annotated in Bengali language by the Bengali speaking people.

### 3.1.2 Encoder

The encoder encodes the image into smaller representations of the original image. We use CNN as an encoder. The ResNet-101 encoder generates and encoding image of size 14x14 with 2048 learned channels. The encoder in composed of the feature extraction using transfer learning, features that are further passed through a fully connected layer that also takes care of the resizing.

### 3.1.3 Decoder

The decoder takes the encoded features from the encoder. The decoder generates a caption word by word from the encoded images. We will use LSTM as a RNN. Weighted encoded images indicates the decoder where should it pay more attention.

### 3.1.4 Attention

Attention network computes the weight of important pixels. The attention mechanism considers the generated sequence and attends to the part of the image that needs describing next.

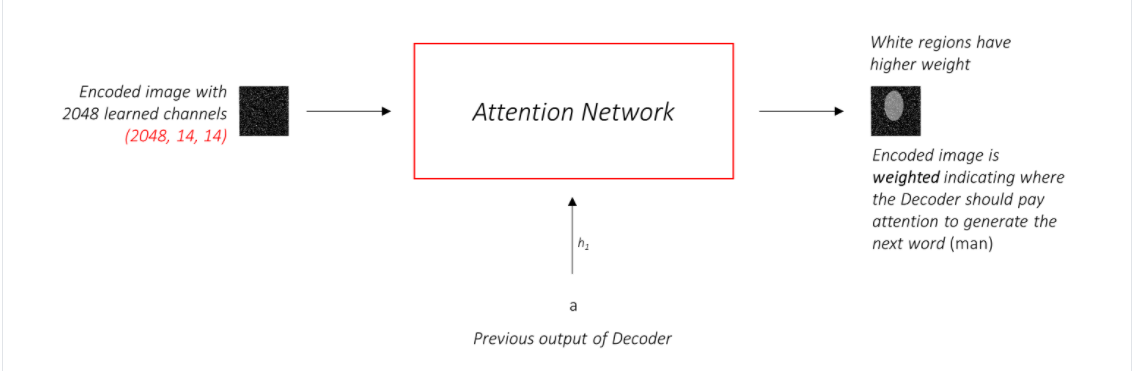


Fig. 1 Attention Network weight to the encoded image

### 3.1.5 Beam Search

To find the best combination of words for an image caption we use beam search. We also use a linear layer to transform the Decoder’s output into a score for each word in the vocabulary. Beam search choose the sequence with the best overall score.

## 3.2 Limitations and Challenges

The system that we want to develop has some advantages over other system of automated Bengali image captioning.

### 3.2.1 Limitations of other system

Oboyob Bengali images captioning system uses the Flickr-8k dataset that has English annotation. Because of the scarcity of the data and resources they take the approach of translating the English captions into Bengali which slows the process. Chittron on the other hand has used the same dataset but cannot generate appropriate caption sometimes. It also sometime mistakes grammatical mistakes.

### 3.2.2 Improvements in our system

We will use the latest dataset of Bengali annotations. We have now the latest library and resources to develop our system. We will use attention network and beam search to generate accurate and right Bengali caption from the images.

# 4 Methodology

Automatically generating natural language sentences describing an image generally has two components:

1. Extracting the visual information,
2. Describing it in a grammatically correct natural language sentence.

## 4.1 Convolutional Neural Network (CNN)

There is a need for a model with a large learning capacity to learn about thousands of objects from a large number of images. Deep learning presents computational models that are composed of multiple processing layers to learn representations of data in images. Deep learning based Convolutional Neural Networks play a key role in many applications, one of which is image recognition. Image recognition is used to perform a large number of visual tasks, such as understanding the content of images. There are several well-known models in the field of CNNs based on object detection and segmentation.

## 4.2 Recurrent Neural Network (RNN)

Sequence models like recurrent neural network (RNN) have widely been utilized in speech recognition, natural language processing, and other areas. Sequence models can address supervised learning problems like machine translation, name entity recognition, DNA sequence analysis, and sentiment classification.

## 4.3 Gated Recurrent Unit (GRU)

Gated recurrent unit (GRU) is a gating mechanism in RNN, introduced in 2014 by Cho et al. The basic RNN algorithm runs into vanishing gradient problem (a difficulty in training artificial neural networks). The gated recurrent units are an effective solution for addressing the vanishing gradient problem. They allow neural networks to capture a much longer-range dependency. The advantage of the GRU is that it is a simple model and so it is actually easy to build a big network. Also, it only has two gates, so it computes quickly.

## 4.4 Long Short-Term Memory (LSTM)

LSTM, as a special RNN structure, has proven to be stable and powerful for modeling long-range dependencies in various studies. LSTM can be adopted as a building block for complex structures. The complex unit in Long Short-Term Memory is called a memory cell. Each memory cell is built around a central linear unit with a fixed self-connection. LSTM is historically proven more powerful and more effective than a regular RNN since it has three gates (forget, update, and output). Long Short-Term Memory recurrent neural networks can be used to generate complex sequences with long-range structure.

## 4.5 Image Captioning Evaluation Metrics

Captions are evaluated using the BLEU, METEOR, and other metrics. These metrics are common for comparing the different image captioning models, and have varying degrees of similarity with human judgment.

**BLEU:** Bilingual Evaluation Understudy is a method of automatic machine translation evaluation that is a precision-based metric, correlates highly with human evaluation, and has a little marginal cost per run BLEU has different n-grams based versions for candidate sentences with respect to the reference sentences.

**METEOR:** Metric for Evaluation of Translation with Explicit Ordering is an automatic metric that evaluates translation hypotheses. It is based on a generalized concept of unigram matching between the machine-produced translation and human-produced reference translations.

## 4.6 Software Requirement

**PyTorch:** PyTorch is a Python-based scientific computing package that serves two purposes: as a replacement for NumPy to use the power of GPUs and as a deep learning research platform that provides maximum flexibility and speed.

**TensorFlow:** TensorFlow is an end-to-end open source platform for machine learning. TensorFlow is developed by Google and has integrated the most common units in deep learning frameworks. It supports many up-to-date networks such as CNN and RNN with different settings.

## 4.7 Deep Learning Models

**Sequence to sequence (Show and Tell):** Uses a multilayered Long Short-Term Memory (LSTM). One LSTM is used to read the input sequence to obtain large fixed-dimensional vector. Another LSTM is used to extract the output sequence from that vector It can learn data with a long-range temporal dependency. The model stops making predictions after it reaches the end of sentence token.

**Attention on Attention (AoA):** An encoder-decoder based model. Image is first encoded via a CNN-based network to a set of feature vectors. Decoded to words via an RNN-based network. Expands the traditional frameworks of attention to assess the connection between the effects of attention and the queries.

**Neural baby talk:** NBT first generates a sentence ‘template’ with slot locations explicitly tied to specific image regions. It explicitly localizes objects in the image while generating free-form natural language descriptions for image captioning. The object detector will detect the region and CNN will analyze the region features. Based on that RNN will generate image caption. So, the slots get filled with proper text based on the image.

# 5 Results and discussion

So far, we have some understanding on various models and how they can be used in captioning an image through reading papers. We are still working on finding the best model that can provide us with the appropriate performance. For training the model, we already have our dataset. To improve our practical knowledge, we tried to run a sample code on Automated image captioning in English Language.

Our goal is to produce Bangla caption when the input is an image. We have faced some errors while running the sample code. We are trying to fix the errors and execute our desirable output. Hope we will find a solution sooner. We are all working patiently to accomplish our aim of making this project successful within our time boundary.

# 6 Conclusions and recommendations

In this project, we proposed to build an automated image captioning system in Bangla language by using deep neural network. We will build such a model which can read an image and generate a caption for that particular image. Our system will correctly predict the contents of a picture and provide an accurate caption. The dataset we are going to use in this project is “BanglaLekha-ImageCaptions” with 16,000 images, this all are collected from public domain and almost all the images are somehow connected or captured in Bangladesh. We find some models that will use to build our project after reviewing some paper on automated image captioning, Convolutional, Recurrent neural network and other related papers. Models like Attention on Attention (AoA), Sequence to Sequence (Show and tell)

And Neural Baby Talk. Show and Tell model uses a multilayered Long Short-Term Memory (LSTM) which is very effective for sequential data. This model analysis will help us to implement our future work. We have also a plan to implement this system as a web application and a mobile application in future.

# 7 Acknowledgements

We gratefully acknowledgement the support of our course instructor Mohammad Ashrafuzzaman Khan. His guidance and suggestions are proved very helpful for us.

# 8 References

This is the last section of the report, prior to any appendices. The references should not be double-spaced, but single-spaced. For a technical report, use the CSE style.

[1] <https://towardsdatascience.com/automatic-image-captioning-with-cnn-rnn-aae3cd442d83>

[2] <https://www.analyticsvidhya.com/blog/2018/04/solving-an-image-captioning-task-using-deep-learning/>

[3] <https://www.hindawi.com/journals/cin/2020/3062706/>

[4] <https://arxiv.org/pdf/1809.00339.pdf>

[5] <https://www.ijain.org/index.php/IJAIN/article/view/499>

[6] <https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning?fbclid=IwAR0fCU8lzclspbDCUubL5eVC7rQjMrAw44NrudQopB_r1GKSibgNBTwFEmE>

[7] <https://content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs179351>

[8] <https://csce.ucmss.com/cr/books/2019/LFS/CSREA2019/IPC3562.pdf>

[9] <https://www.researchgate.net/publication/13853244_Long_Short-term_Memory/link/5700e75608aea6b7746a0624/download>

[10] <https://arxiv.org/abs/1612.07600>

[11] <https://openaccess.thecvf.com/content_cvpr_2015/html/Donahue_Long-Term_Recurrent_Convolutional_2015_CVPR_paper.html>

[12] <https://www.academia.edu/43726642/Object_Detection_using_Deep_Learning_Algorithm_CNN>