Image Caption Generation

Course Code: CSE 4120
Course Title: Technical Writing & Seminar
Date: 2 June,2024

Department of Computer Science and Engineering
Khulna University of Engineering & Technology, Khulna

Supervised by:
Dr. K.M Azharul Hasan
Professor of Department of
Computer Science and Engineering
&
Sunanda Das
Assistant Professor of
Department of
Computer Science and Engineering

Submitted by:
Sabnaj Akter
Roll: 1907042
Year: 4th
Semester: 1st
Department of Computer Science and Engineering
Khulna University of Engineering & Technology

Contents

List of Figures

List of Tables

1 Abstract

Image captioning is an innovative technique that entails generating sentence-like constructs to describe a scene depicted as an image. This specific type of algorithm has the intention of identifying a few entities on the image, do a couple of operations and try to detect a few characteristics of the image in question. When the system has got this information, it should be able to farther advance and provide most appropriate and precise description of the image under context and there should not be any syntactical/semantic mistake. Probably, with growth of learning progress the algorithms write the text in natural appearance as it is done attempting to describe an image in a set of natural sounding sentences. Of understanding and generating descriptions of the contents in an image, this is an inherent ability in humans, and is still a challenging task when handled by machines at the present. Thus, enumerating all possible applications, we remain on the level of a brief overview of numerous applications of named entity recognition. The work involves the generation of 8 short descriptions based on practical skills such as Natural language processing (NLP), Computer vision (CV), and Deep Learning (DL). The generated captions are also described in these papers, highlighting the attention technique and the encoder-decoder. CNN refers to Convolutional Neural Network in this context, which we have used here in the form of its pre-trained model. Not really possible to get what object is depicted on the image, but it is possible to get the features of the image through Inception V3 and then use the Recurrent Neural Network (RNN). Here, the GRU will give the opportunity to generate a relevant caption for this process. Within the processes organized to design the captions, the listed models focus on an attention mechanism referred to as Bahdanau attention. As for the method of training the model, one has the MS COCO dataset and the Flickr8k corpus at one's disposal. They agree to the fact that the model is inclined quite effectively at the intended purpose of interpreting images and translating them to texts.

2 Introduction

Image caption generation is a famous subject amongst researchers and a lot of work has been done in this field in the recent past. Many techniques were produced in order to generate accurate captions but not all were as efficient as the methods which are available today. Today the efficiency and accuracy has enhanced by great deals. With the passage of time, the researchers have come up with lots of new methodologies for classifying images. Deep Learning is a widely used concept and is almost a recommendation for every classification. Neural Network, convolutional neural network, recurrent neural network, are some of the well known deep learning techniques widely used for caption generation. The main purpose of this system is to generate image captions given as input by the users. This will help in organizing and labelling images in efficient manner without human involvement

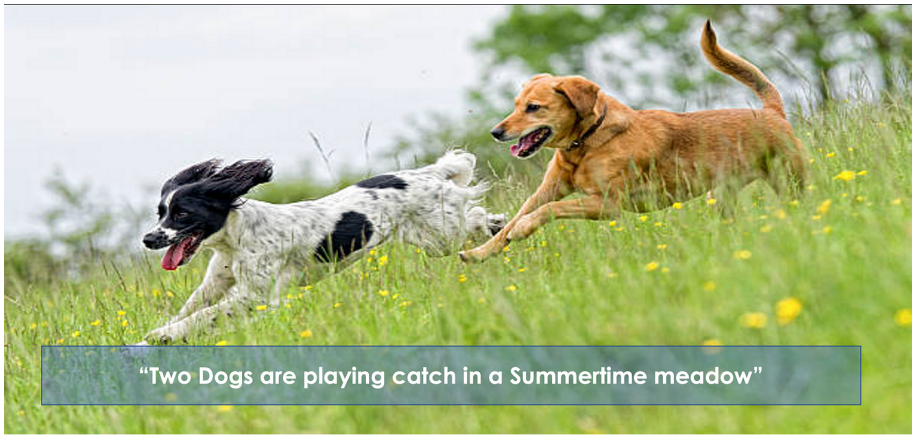

Figure 1. Caption Generation

By generating captions for multiple images of a same file one can organize or arrange those files easily and quickly. The people who are blind or the ones who have low vision can understand the images by their caption or description provided by the image captioning process. The images added in a website could be understood well when it has a valid description to it. Therefore, the process of generating a website could be done quickly by just adding the images and the caption for it could be found by applying the image caption generator. Hence, image captioning has become a critical tool for web development

3 Background and Problem Statement

3.1 Traditional Approaches

Early attempts at image captioning relied heavily on template-based methods and retrieval-based methods:In previous stages of Developmental image captioning methodologies, some of the initial methods deployed were template-based methods and strategies and retrieval-based methods and strategies:

Template-Based Methods: These approaches are based upon reading through a number of templates that are placed in various categorizations including objects, activities, and qualities of the image that were discovered. It is quite easy to implement these procedures, but they come with the strict templates of the rule books that need to encompass the entire procedure.

Retrieval-Based Methods: These methods include search an image in the input image way for selecting images from a large number of images and Using the captions of the selected images. The advantages of the method, namely, the transformation of images, the ability to generate captions for images, and recognition, are less valuable due to the drawbacks; it cannot create captions for unknown images and is less flexible than the previous method and necessitates other similar images.

3.2 Machine Learning Approaches

With the advent of machine learning, more sophisticated methods began to emerge, particularly those leveraging feature extraction and probabilistic models:With the advent of machine learning, more sophisticated methods began to emerge, particularly those leveraging feature extraction and probabilistic models:

Bag of Words (BoW): Images are represented as an entity of visual features that may be provided with the help of techniques such as Scale-Invariant Feature Transform (SIFT). These features are then converted to the words using machine learning techniques such as; support vector machines (SVM) or decision trees. But this kind of approach has a failure in modeling the temporal relationship of language.

Probabilistic Models:HMM and CRF were used to create augmented captions of the images, given the probability of sequences of words which was used given the extracted features from the images. These models gave a probabilistic view but had several drawbacks that were associated with the dependence on features which were hand crafted and the inability for these models to capture high level features in database like information.

3.3 Deep Learning Approaches

The advancement that led the way to generating captions for images happened at the time when deep learning approached the problem with Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Such end to end trainable systems were introduced during this era that made it possible to learn from raw data.

Convolutional Neural Networks (CNNs):CNNs are seen as tools capable of generating multiple representations of images, from more simple to much more complicated. AlexNet, VGGNet, ResNet have tremendously improved the convolutional layer where the feature extraction of the images which is very important in the development of image captioning systems. Recurrent Neural Networks RNNs:Are a class of neural In which a wide variety of topics is being discussed memetic algorithms that deal with partial differential equations and other problems.It tried to address the needs of networks that are specialized in sequential data processing.

Recurrent Neural Networks RNNs: Unlike other neural networks, RNNs have feedback which allows the sequence of connections to form recurrent loops which provide a 'memory' for previous inputs. This makes them suitable for uses where the context and/or the sequence of the inputs and outputs are important for the application such as language modeling speech recognition and time series data processing. In image captioning, RNNs, also LSTMs, usually combined with CNNs, accept the vectors of the image features which are created by CNNs, and provide accurate and meaningful descriptions which are based

on the context by analyzing the chain of letters of the alphabet and forming words in connection with other words.

4 Review of Literature

CNN(Convolutional neural network),RNN(recurrent neural network),LSTM(long short-term memory networks),Attention Machanism are used to generate caption from an image.

Title	**Used Method**	**Publisher & Published Year**
Neural Image Caption Generation with Visual Attention	Visualization	CSREA Press-2015
Image caption generation with high-level image features	Attention Mechanism	IEEE-2019
Generating Image Captions using Deep Learning and Natural Language Processing	CNN and RNN	IEEE-2021
Image Caption Generator using Attention Mechanism	CNN,RNN and Attention Mechanism	IEEE-2021

Table 1: Literature Review

4.1 CNN

Convolutional Neural Network (CNN) the type of Deep Learning algorithm which is developed to work with networks of hooked matrices such as images. It is made up of convolution layers that are utilized in extracting filters from data that is fed into it so as to produce spatial hierarchies of the features present in the data. These are followed by the pooling layers that reduce dimensionality keeping maximum information and reducing computational processing. CNNs are especially suitable in recognizing features of several hierarchical levels, which including edges, textures and shapes. Some of the most widespread practices of CNNs usage are image classification, object recognition and image annotation. Speaking about the work on this type of neural networks, it should be noted that such neural networks are constructed in a way that features existing in the image under analysis are selected and extracted by Neural networks of the given models are more accurate and work more quickly than traditional methods.

4.2 RNN

The RNN is a specific type of the neural networks which is designed to work with temporal sequences. However, unlike the normal neural networks circuits, the RNNs include 'feedback connections' where a node connects back to the previous state which makes the circuits able to remember previous value of the input. Hence it can be employed on environments that require not only the context but also the positioning such as Word2Vec, Sphinx for speech recognition, and time series prediction. Indeed in image captioning, the RNNs that are used, and which are called Long Short-Term Memory (LSTM) networks, work on the basis of words and potential dependences, and receive as input the feature vectors linked to the topology at the level of the CNNs.

4.3 LSTM

LSTM is one of the RNN models which is built solely for temporal modelling or to handle

temporal dependencies and in order to overcome the vanishing gradient problem. They arrive at this stage by use of memory cells and gate control whereby they are able to sort out information that is relevant. LSTMs are particularly suitable for instances where sequences in natural language or any streams of events are required such as language modeling, translation, and image captioning in sequence, hence, LSTMs generate text that is contextually relevant, to the inputs streams.

4.4 Attention Mechanism

The attention mechanism is a deep learning mechanism that is used in neural networks to amplify the results of models during training by directing the model's focus to the critical information within the input data. This results in the model evaluating elements of varying significance differently, helping it identify which information is crucial. This approach is considered very useful in tasks such as MT and IC since it enhances performance by altering its emphasis in accordance with the state.

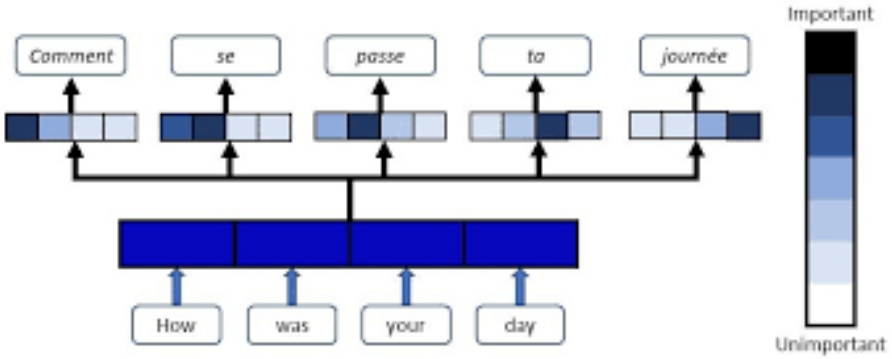

Figure 2. Attention Machine

5 Methodology

5.1 Generating Image Captions using Deep Learning and Natural Language Processing

5.1.1 System Design

This framework is used to develop a model, labels for the images according to the picture present in the images. The basic flow illustration of the proposed system is constructed further all the components of this diagram are analyzed properly.

5.1.2 Block Diagram of Proposed System

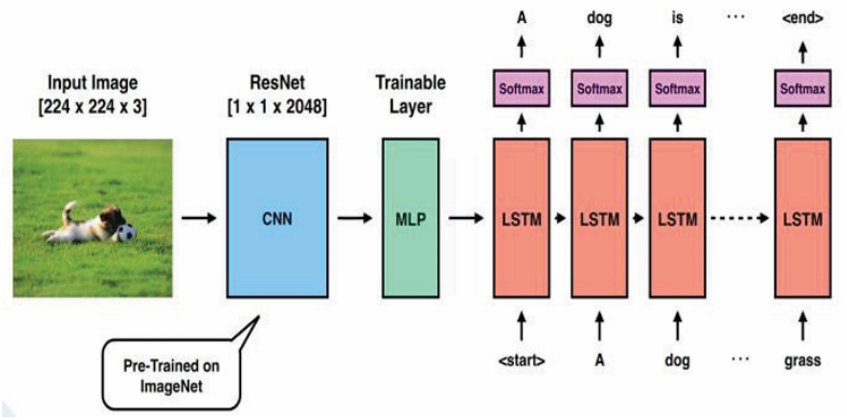

Figure 3. Diagrammatic representation of the proposed model

5.1.3 Workflow of the Proposed System

Here the intended idea is to engender captions or descriptions for diverse images used as sample inputs. For accomplishing this, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) are the two types of deep learning algorithms are implemented. The image or picture gets passed through CNN to recognize objects and scenario present in a particular image. Through CNN, few more concepts were applied like pooling, padding and use of filters, etc. Basically, CNN is used to bring out all the features present in a sample image. Additionally, in transfer learning where two main projects are considered: Glove and Inception V3. Glove defines the set of NLP vectors for common words and Inception V3 is used for bringing out important features from the image.Further, through Natural Language Processing (NLP) the images are processed to simpler formats which makes easier for us to communicate with the computers.

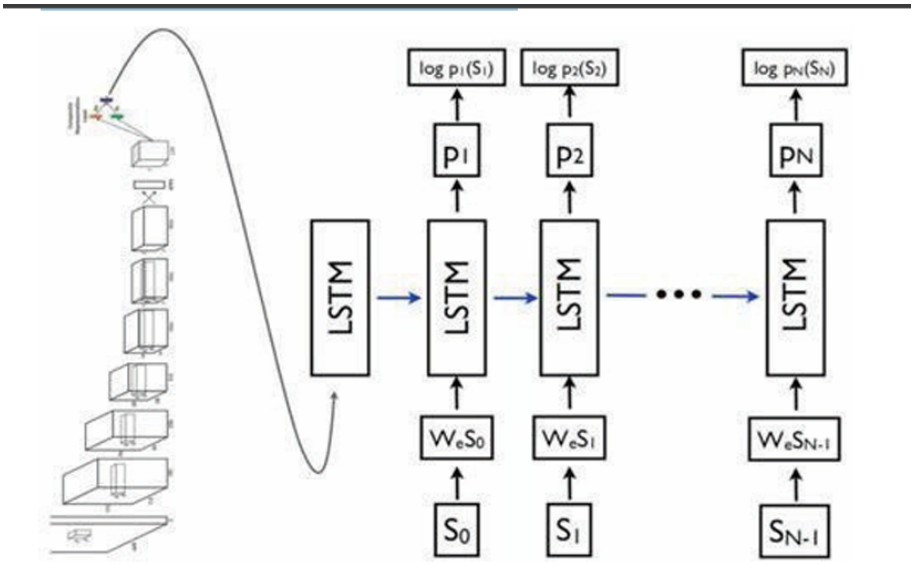

Figure 4. Basic workflow of the model

5.2 Image Caption Generation using Deep Learning Technique

In this work, neural framework is proposed for generating captions from images which are basically derived from probability theory. By using a powerful mathematical model, it is possible to achieve better results, which maximizes the probability of the correct translation for both inference and training.

5.2.1 Convolutional Neural Network (CNN)

The convolutional networks are currently used in visual recognition. There are number of convolutional layers in CNN. After these convolutional layers, next layers are fully connected layers as in multilayer neural network [14]. The CNNis designed in such a way that the benefit of 2D structure of input image can be taken. This target is accomplish with the help of number of local connections and tied weights along with various pooling techniques which result in translation invariant features. The main advantages of using CNN are ease of training and possessing less parameters as compared to other networks with equal number of hidden statesss. For this work, we are using Visual Group Geometry(VGG) network, which is Deep CNN for large scale image recognition [15]. It is available in 16 layers as well as 19 layers. The classification error results for both 16 and 19 layers are almost same for validation set as well as test set, which is around 7.4and 7.3%. This model gives the features of images which are used in further process of caption generation.

5.2.2 Long Short-Term Memory (LSTM)

The transitory dynamics in a set of things are modelled by using a recurrent neural network [17]. It is very difficult for ordinary RNN to acquire long term dynamics as they get vanished and exploding weights or gradients [9]. The memory cell is main block of LSTM. It stores the present value for long period of time. Gates are there for controlling update time of state of cell. The number of connections between memory cell and gates represent variants. Our model is based on the LSTM block which depends on the LSTM with no peephole architecture .

5.3 Image caption generation using Attention Mechanism

To implement this, the model contains two systems. The first converts the image information into features, and the second converts the features into a meaningful English sentence (caption). The first part can be called the encoder, as it encodes image information into a feature vector. The second part will be referred to as a decoder that converts the features into a caption. Fig. 1. represents this classic image caption generator flowchart.

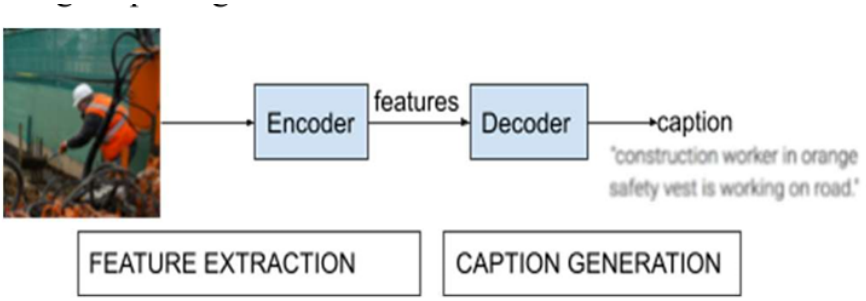

Figure 5. Flowchart for classic image caption generator

5.3.1 Model Overview

Speacified Algorithms which have been used in the development of the encoder comprises of of: Convolutional Neural Networks(CNN) should not be used anymore which were removed in images and it was designed to extract necessary features from the images.The CNN used in this study is titled the : The InceptionV3 model is imported from the keras library and the model is instantiated in pre-trained one encoding part. Application A Bahdanau attention it is marked as attentionus and at each Step of each attention iteration is concat is used to join all matrices and all of which have sizes equal to the number of matrices joined along the first axis.Specifically, in terms of processing a sequence of elements,it uses several mechanisms. Another approach used in our model is Recurrent Neural Network (RNN) for sequence modelling for the decoder part. About the kind of the gradient-based neural network applied for this investigation, it is called as the Gated Recurrent Unit (GRU) .

5.3.2 Convolutional Neural Network

Inception V3 was built for object detection. Upon consuming an input image of (299x299x3), it outputs object wise probabilities. Since InceptionV3 is an object-detection deep network, it needs to be tweaked a little to work as an encoder. To get a feature vector from this deep network, the final layer that classifies the images into objects is removed. This layer is a fully connected layer that classifies and predicts the objects. Once the fully connected layer on top is removed, a feature vector of size (8x8x2048) is obtained from InceptionV3. The internal architecture of this encoder (InceptionV3) is summarized in the table below.

5.3.3 Recurrent Neural Network

The decoder mechanism is another RNN that generates words in natural language. The Recurrent Neural Networks/ RNNs some of the common ones are Vanilla RNN, Long Sho.However, there are still many others such as Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), and others. Vanilla RNN is absolutely disapproved as it results to the vanishing gradient issue we faced, that is, the words that are predicted first, are no longer accessible to the consciousness when it predicts the next words, where the word “A1” refers to the actual next word in the text and “A2” refers to the predicted next word by the system. Hence the first words, at last, become very insignificant in the sense that even their weights themselves are extremely low in the computed output.In order to resolve this problem, LSTM (Long Short Term Memory) was introduced. Although it is slightly less complicated than LSTM, GRU is almost as powerful as it in terms of performance complex in cell architecture. Thus, the focus was put on utilizing GRU to implement this caption generator.

6 Result

Field	**Generating Image Captions using Deep Learning and Natural Language Processing**	**Image Caption Generation using Deep Learning Technique**	**Image caption generation using Attention Mechanism**
Result Analysis	No specific metrics mentioned, likely uses standard metrics like BLEU	BLEU scores: Flickr8k = 0.53356, Flickr30k = 0.61433, MSCOCO = 0.67257	Attention mechanism aims for better accuracy; specific metrics not provided.
Dataset Analysis	Flickr8k dataset, which contains 8,000 images with captions.	Three datasets: Flickr8k, Flickr30k, and MSCOCO.	MSCOCO dataset with 82,000 images, each having five annotated captions.
Error Analysi	Errors in generated captions due to misidentification of objects and scenes.	Errors include descriptions with minor inaccuracies.	Errors occur when the attention mechanism misidentifies important image regions.
Limitation	Limited dataset size restricts the model's ability to generalize well.	Model performance varies with dataset size.	The model's effectiveness is heavily dependent on the attention mechanism.

Table 2: Result Analysis

7 Finding and Recommendation

7.1 Finding

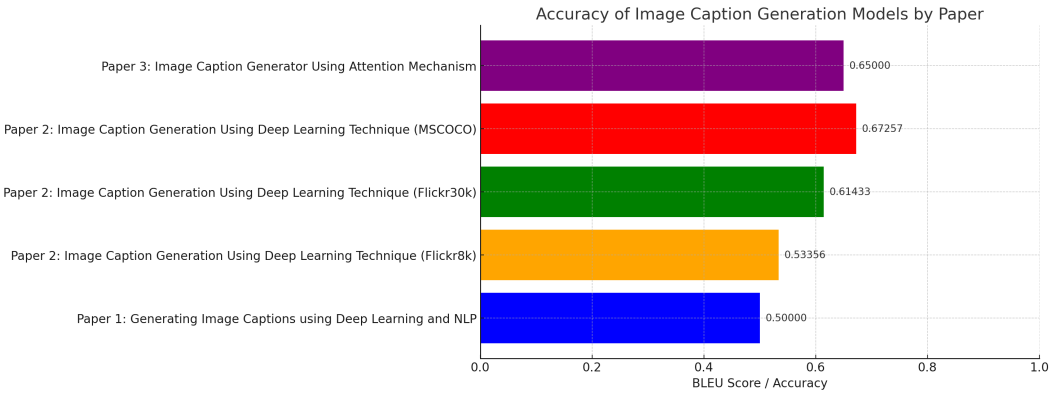

Figure 6. Accuracy of Model

7.2 Recommendation

According to the results depicted in the diagrammatic representation of the proposed model (Figure 6), it is suggested to integrate an aspect of Deep Learning categorized as DL trio with an attention mechanism for image caption generation, integrated with NLP. The basis for this shall be founded on several observations of the interdependent benefits associated with each component in facilitating the overall accuracy and aptitude of the captions that are generated. This paper focuses on the combination of Deep Learning and Natural Language Processing to enhance the process and understanding of writing image captions. CNNs are very effective at extracting finer details in images making DL the best choice for images, whereas RNNs and Transformers make coherent and contextually relevant descriptions possible for NLP techniques.

8 Addressing Course outcomes and program outcomes

This course is intended to teach the necessary skills to the students to tackle complicated engineering problems systematically This course is aimed at providing the students with structured methodology to solve multi faceted engineering problems. Students will also be able to design complex engineering problems, to define them, and to solve them from the first principles of mathematics, science and engineering sciences to ensure that they develop broad problem solving skills.With increased focus on the professional standards of behavior, the course fosters an understanding of the purpose of ethics and the role of the professional, ensuring students understand their responsibilities within the professional world.

Also, it introduces the philosophy and concept of independent and team-work in the course work. Students will be able to understand how to work as an individual and as a team, as well as how to work in groups in a successful multi-disciplinary and/or cross-functional team or leading one, which underlines the key aspects of both teamwork and leadership to solve engineering problems.Promoting students' communication is also another important factor: they are to describe complex engineering processes verbally. It may involve stating reports, developing design documentation, and preparing and

presenting presentations, as well as giving and receiving clear instructions which are crucial for career achievement and social life.

9 Addressing Complex Engineering Activities

Sustained engineering practices, undertake tasks that could be described as complex due to their diverse and wide-ranging nature. This involves a creative utilization of the productive fundamental knowledge as well as engineering particulars into innovative undertakings. These disasters usually come with a significant impact on the population and the ecosystem, and thus, their prediction and prevention can be a difficult feat. Furthermore, they go beyond prior experiences since they employ principles-based means, which, in a way, expand the know-how and experience frontiers.

To sum up, it is possible to underline that there are a number of features that are characteristic of complex engineering activities, namely, the application of engineering concepts in creative manner, high degree of public and environmental concerns in terms of engineering projects, and the potential for exploring new opportunities and challenging one engineering forward in new and sophisticated ways. These activities involve engineers in managing technical, societal, and environmental demands as they strive to develop their plans that conform to professional standards in engineering.

10 Conclusion

Inconclusion, image caption generation has become an imperative asset with multifunctional usage in multiple subjects. It benefits organization and improves the content's access and understanding for people with a visual impairment; due to it, the task of creating descriptive captions for images is most efficiently executed with the help of specialized software. Indeed, the neural network models introduced here, especially Incorporating encoder-decoder networks with attention, reveal respective improvements. These models efficiently work for extracting features from the images or for image descriptions than the conventional techniques and provide more meaningful and logical captions. However, issues are still open, for instance, what to do with elements which are not determined by the model, in other words, the objects are out of the model's vision, the captions may not be so valuable or even misleading. The limitations of the current approaches must be solved in the future research; several ways can be suggested: increase in the size of the training corpus, appeals to unsupervised training to enhance the accuracy of the achieved results. In conclusion, with the advancements that have been seen in the field of image captioning it remains an essential part of image processing with potential and continually expanding utilization in the future for numerous areas and applications in both industrial and academic sectors.

11 References

1. Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019), "A comprehensive survey of deep learning for image captioning", ACM Computing Surveys (CSUR), 51(6), 1-36.

2. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2016). "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge", IEEE transactions on pattern analysis and machine intelligence, 39(4), 652-663.

3. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., & Berg, T. L. (2013). "Baby talk: Understanding and generating simple image descriptions", IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12), 2891-2903.

4. Farhadi, Ali, et al. "Every picture tells a story: Generating sentences from images." European conference on computer vision. Springer, Berlin, Heidelberg, 2010.

5. Yang, Yezhou, et al. "Corpus-guided sentence generation of natural images." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.

6. Kulkarni, Girish, et al. "Babytalk: Understanding and generating simple image descriptions." IEEE Transactions on Pattern Analysis and Machine Intelligence 35.12 (2013): 2891-2903.

7. Moolayil, Jojo John. Learn Keras for Deep Neural Networks: A Fast Track Approach to Modern Deep Learning with Python, 1st edn, Apress, New York, 2019.

8. Gupta, Shailja, et al. "Application-based attention mechanisms in deep learning." International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) ICRITO, September 2020.

9. Chohan, Murk, et al. "Image captioning using deep learning: a systematic literature review." International Journal of Advanced Computer Science and Applications, Vol. 11, No. 5, 2020.

10. Yu, Yong, et al. "A review of recurrent neural networks: LSTM cells and network architectures." Neural Computation, Volume 31, Issue 7, July 2019.

12 Publication Details

SL NO	TITLE	AUTHORS	CONFERENCE/JOURNAL NAME	Publisher & Published Year
1	Generating Image Captions using Deep Learning and Natural Language Processing	Subhash Chand Gupta,Akshita Tyagi,Tulsi Sharma ,Rana Majumdar	2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) Amity University, Noida, India. Sep 3-4, 2021)	IEEE-2021
2	Image Caption Generation using Deep Learning Technique	Chetan Amritkar,Vaishali Jabade	2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)	IEEE-2018
3	Image caption generation using Attention Mechanism	Vaishnavi Agrawal ,Neha Tuniya ,Vibha Vyas, Shariva Dhekane	212th ICCCNT 2021 July 6-8, 2021 - IIT - Kharagpur Kharagpur, India	IEEE-2021

Table 3: Publication Details of the Papers