International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Image Captioning using Deep Learning Model for Visually Impaired People

Shailesh Shivaji Patil[1], Premal J. Patel[2]

[1,2]Ipcowala Institute of Engineering and Technology (IIET), Gujarat Technological University, Dharmaj, Gujarat, India

Abstract
The process of generating meaningful textual description for an image is known as image captioning. The ideal caption for the image not only includes object and their characteristics but also emphasizes the action performed by the objects. In image captioning, there are two main tasks. The first crucial task involves effectively recognizing objects present in the given image. Once all the objects are identified along with their characteristics, the dense model is to identify the correct action or verbs associated with the recognized objects. In image captioning, the second part involves creating the subsequent phase that connects all the recognized objects with their respective attributes and action. This paper focuses on generating captions for images using Deep Learning for Visually Impaired People.

Keywords: Deep Learning, CNN, RNN, LSTM, Image Processing.

1. Introduction

The saying of "One picture is worth a thousand words an interface worth a thousand picture" is well known to everyone. Captioning an image can be done in numerous ways but the determining most appropriate caption for an image is most challenging task. Many surveys had been conducted on the topic Image captioning for identifying the best caption generation model. In two primary groups Image captioning model can be categorized. One of the most widely using categories is supervised learning. In this category, the training images mainly come with the label and these labels assist in generating captions for the test images for utilizing input and output pairs. However, the drawback of employing the supervised learning approach is that the model may fail to identify new objects that are absent for the training data set. The second category that overcome the disadvantages mentioned in the supervised learning. The category learns from unlabeled test data [1].

Image captioning models commonly adopt an encoder-decoder architecture, utilizing abstract image feature vectors as input for encoder to generate caption. Creating a natural language description from images is significant challenge within the intersection of computer vision, image processing, captioning Image, natural language processing and artificial intelligence which generating natural language descriptions automatically based on the observed content in an image, plays a crucial role in enhancing scene understanding, this integration involve leveraging the expert knowledge from both computer vision and natural language processing [2].

A traditional algorithm, combining Convolutional and Recurrent network used for generating captions, faces various issues, including gradient vanishing, imprecise identification of objects and their relationship, and generation of captions solely for familiar images. A variation of the traditional method, the automatic image captioning model combines advanced Convolutional and Long Short-Term Memory Deep Neural Network algorithms (CNN and LSTM) to overcome the issue that arise with the traditional way of captioning. Divided into two stages, model employs the Convolutional algorithm in first stage and Long Short-Term Memory in second stage. The image/picture serves as the input to the first stage. The proposed system model emphasizes generating informative captions that best describes the image scene [3]. Simple flow diagram of image captioning is shown in Figure 1.

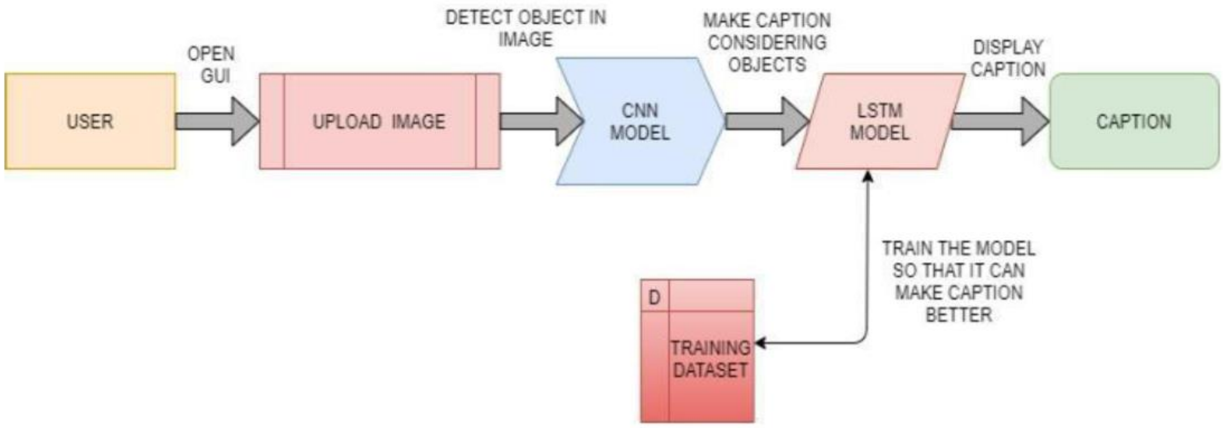

Figure 1: Flow Diagram of System

2. Literature Review

No.	Author(s)	Approach	Description	Limitation
1	Chetan Amritkar, et. al., 2018 [4]	CNN And RNN	This model incorporates both recurrent neural network (RNN) and Convolutional Neural Network (CNN) are used to extract feature from image. The model is trained so that when an input image is provided as input it produce the captions which clearly describes the image.	The model's descriptions or captions are divided into three categories: description error free, description containing few small mistake, image and description are slightly linked but, not at all. The categories in the results are caused by the proximity of some specific words; for example, when a 'vehicle' is nearby, term like "vehicle"," Car"," Van "etc. are also formatted, which may be inaccurate it is evident from a vast number of studies that using larger dataset improves the model's performance.

				Both accuracy and losses will decrease with the greater dataset. It is also be fascinating to see how unsupervised data for text and images may be utilized to enhanced methods for creating a caption for image.
2	Jyun-You Lin, et. al., 2020 [5]	YOLO, CNN, LSTM	They suggested a smart glasses system which is Based on deep learning system for visually impaired people. By capturing photo from the camera function of the smart glasses, the system can upload captured images to our object detection system which function at backend and provide voice speech of caption helpful to understand visually disabled individuals regarding the object Infront of them.	The suggested method requires in 3.788 seconds to uploading photos and generating voice results with 96.3% success rate in object detection. In order to enhance the quality of for visually impaired individual with expect that our application will help them to comprehending their surrounding and engage in closer social interaction.
3	Boeun Kim, et. al., 2019 [6]	CAM	They presented an approach for generating multiple captions using a variation auto encoder, a type generative model. Give the significance of image features in caption generation, a proposed method involves extracting Caption Attention Map (CAM) to highlight key areas in the image.	A network is proposed that varying the guide map reflecting the style associated to region of image, creates multiple captions. The CAM vector creates latent space which we first receive at VAE encoder the Vector representing the image attention region style is then taken out of the laten space. This vector serves as condition to construct the caption at the VAE decoder where it is used at

				the CNN input. In terms of variety, the proposed model out forms the basic model with similar accuracy. We verified via the use sample cases, that the caption produces by the suggested method have a higher diversity of expressions and content than those produce by the basic model. The suggested approach might be helpful.
4	Genc Hoxha, et. al., 2019 [7]	RNN and CNN	The RS image retrieval system have three primary processes. The first one involves generating the Textual description with Convolution neural network (CNN) and produces written explanations of the images' contents. In order to extract features from images, a convolutional neural network (CNN) and a recurrent neural network (RNN) are combined in the second step. The third step involves producing the descriptions and details of the content, respectively.	With the goal of examining the high-level sematic material buried in the created descriptions, we have presented in this study a sematic picture retrieval technique based on generated textual description. we find that there is an average difference of 0.3 in mean BLEU score when comparing the produced description and the genuine description for RS pictural retrieval. Our goal is future work is to enhance the caption generating block in order to close this gap and boost retrieval capabilities
5	Simao Herdade, et. al., 2019 [8]	ORT	In this study they introduce the Object Relation Transformer, which extended the methodology that builds the approach by explicitly adding the spatial relationship information	Currently only geometric information is considered by our model at the encoder stage. Next, we plan to add geometric attention to the layer of decoder cross attention between words and

			between input identified objects through geometric attention	objects. Our goal is to do this by clearly linking object bounding boxes with decoded phrases. This should result in increased a performance gains and better model interpretability
6	Shuang Liu, et. al., 2018 [9]	CNN and RNN	Image captioning represent this field, where computers are trained to comprehend the visual content and generated descripted caption of an image using one or more sentence. The process of generating meaningful descriptions for high level image semantics is the study of object recognizing from the images to analysing the relationship, state, attributes and characteristics of the objects.	Enhancing the efficiency of content-based image retrieval, describing images through text broadens the application scope of visual understanding in various fields like medicine, security and military, this establishes broad potential application domain. Theoretical framework and research method Simultaneously applied in image captioning can contribute to advancing the theory and application of image annotation and visual question answering (VQA), cross media retrieval, video dialog and video captioning, this has substantial academic and practical application value.
7	Chaoyang Wang, et. al., 2021 [10]	MLP, LSTM	This paper delves into and scrutinizes pertinent research on image captioning. In the beginning, the document introduces the task and potential application contexts related with image captioning. Following that, it analyses	The study examines the fundamental of the few popular image captioning techniques. Introduce as the evaluation indices and data set required in this subject. While the prediction impact has been somewhat enhanced by the current image

			both the encoder decoder structure-based image captioning algorithm. The discussion includes the strength and limitations of each method.	captioning algorithm, they are unable to fulfil the purpose of producing particular description statements based on certain circumstance.
8	Lakshminarasimhan Srinivasan, et. al., 2018 [11]	CNN and LSTM	The authors suggest a hybrid system that make use of Long Short-Term Memory (LSTM) to construct the coherent sentences using the generated keyword and Multilayer Convolutional Neural Network (CNN) to generate vocabulary characterizing the images. After comparing the target image to a vast collection of training photos, the convolution neural network uses the caption it has learned to produce an accurate description.	The creators of this work have used deep learning to provide captions for the photographs. The deep learning architecture was implemented using TensorFlow as a backend and Kera's sequential API. An effective BLEU score of 0. 683 is obtained by the model. The statistic known as BLUE, or Bilingual Evaluation Understudy Score, compares a generated sentence to a reference sentence for assessment. A score of 1.0 is awarded for a perfect match and 0.0 for a perfect mismatch. To enhance the model's feature extraction in the future, the authors are focusing on alternating Pre-Trained Photo Models. Additionally, by applying word vectors to a much broader corpus of data, including news articles and other internet data sources, the authors hope to increase performance even further. While the model's configuration was optimized, it is possible to train alternative

				configurations to observe whether they enhance the picture captioning model's performance.
9	MD. Zakir Hossain, et. al., 2018 [12]	Reinforcement learning, Neural Network	In this survey paper, objective is to provide a through overview of current deep learning approaches based on captioning images and examine the foundational method we evaluate their drawbacks and effectiveness. We additionally delve on the datasets and commonly utilized evaluation metrics in deep learning based automatic image captioning.	This paper, provides a review of methods image captioning based on deep learning. We present a taxonomy of technique for image captioning illustrated a generic block diagram of the major groups, and their advantages and disadvantages. Additionally, we explored various evaluation metrics and datasets, outing their strengths and weaknesses. The paper includes concise summary of experimental results, and we provide a brief overview of potential research directions in this domain. Despite the notable progress made in recent years by deep learning-based image captioning methods, achieving a robust method capable of generating high quality captions for all most images remains a challenge. The continuous emergence of novel deep learning network architectures ensure that automatic image captioning will remain an active research area for some time.

10	Haoran Wang, et. al., 2020 [13]	Combine CNN and kNN, RNN	This paper provides a summary of the pertinent methods which concrete focus on attention mechanism, a crucial element in computer vision in image caption generation has recently become a significant task.	The authors also intend to use word vectors on a much broader corpus of data, including news articles and other internet data sources, in order to increase performance. Although the model's configuration was adjusted, different configurations could be trained to determine if the image captioning model performs better.
11	Raimonda Staniūtė, et. Al., 2019 [14]	NLP, LSTM	In this study, a through SLR (Systematic Literature Review) offers a concise summary of advancements in image captioning over the past four years. The paper main goal is to summarize the research article to finding and describing the most popular method in image captioning along with highlighting the need to raise awareness about incomplete data collection in the paper.	This systematic literature review compiles the most recent papers and their findings in one location to avoid the loss of important concepts and to promote fair competition among the recently developed models. Moreover, there is still uncertainty over the suitability of the MSCOCO and Flick30k datasets for model evaluation, as well as their performance in a variety of contexts. Data will keep growing in volume, and fresh information will keep coming out on a regular basis. Future studies should complete whether static models are adequate for long term applications or there should be increasing emphasis on lifelong learning. We anticipate that this This Systematic literature Rivew (SLR)

				will function as guide and encourage other scientists to collect the latest information for their research evaluation.
12	Syed Haseeb, et. al., 2019 [15]	Deep Learning, CNN, Rest net	In this project, a generative model employing a deep recurrent architecture is utilized, merging recent advancements in machine translation and computer vision the natural sentences that describe an image are produce by the recurrent architecture.	A complete neural network system that can automatically analyse an image and produce a coherent English description. It is predicated on a convolution neural network that compresses an image into a compact representation and then generates a matching text using a recurrent neural network. After receiving the image, the model is trained to maximize the likelihood of the text. Experiments carried out on a variety of datasets demonstrate the robustness in terms of qualitative outcomes (the generated sentences are quite reasonable) as well as quantitative assessments. To evaluate the quality of created phrases, the assessment use ranking measures, also known as BLEU, which is a metric frequently used in machine translation. These studies demonstrate that the suggested approach will perform better as the size of the datasets available for picture description grows.

				In addition, it will be intriguing to observe how image description techniques might be enhanced by unsupervised data, including textual and image-only data.
13	Yang Feng, et. al., 2018 [16]	CNN, RNN	This paper makes the initial effort to train an image captioning model in an unsupervised manner. Our proposed model eliminates the need manually labelled image-sentence pairs, relying instead on an image set, need a corpus of sentence, a set of images, and an established a visual concept decoder. The captioning model learns how to produce	This study abstracts from the usage of paired image-sentence data and presents a novel way to train an image captioning model in an unsupervised way. To the best of our knowledge, this is the initial effort to look into this issue. Our three training objectives aim to accomplish the following: 1) The generated captions should be nearly identical to sentences in the corpus; 2) The image captioning model should convey the object information in the image; and 3) The features of the image and sentence should be aligned in the common latent space and perform bi-directional reconstructions from each other. To aid in the unsupervised picture captioning technique, a massive corpus of over two million phrases describing images was further gathered from Shutterstock. Without utilizing any tagged image-sentence combinations, the experimental findings

				show that the suggested strategy can yield some very promising outcomes. Human assessments for unsupervised picture captioning will be carried out in the future.
14	Moses Soh, 2016 [17]	CNN-LSTM	This work created a generative CNN-LSTM model that approaches by (3.8 CIDEr points lower) the current state of the art while surpassing human baseline by 2.7 BLUE 4 points. The majority of time, experiments on the MSCOCO dataset set produce caption, and we may mitigate the consequence of overfitting by hyperparameter optimizing with dropout and amount of LSTM layers.	With a maintain probability of 75% for dropout and two layers for our decoder LSTM network, we achieved results that are 3.8 CIDEr points and 3.3 BLEU-4 points behind the state-of-the-art after doing a thorough hyperparameter search over the CNN-LSTM model architecture. A comprehensive model search was carried out. Any number of images from the MSCOCO collection can be coherently captioned by a thorough quantitative and qualitative analysis methodology. When people fail to pay close attention to certain characteristics in photos, they often make partial mistakes (e.g., misidentifying a photo of elephants strolling in an enclosure as "elephants in a field"). This shows that this task could be improved by the attention strategies investigated in recent research.
15	Sukriti Rampal, et. al., 2018 [18]	CNN RNN and LSTM	Authors have incorporated an optimized CNN based	In this work, we propose an optimized model for

			encoder, our image caption generator now uses an optimized CNN-based encoder and RNN based decoder model called IndRNN which is more effective at learning longer term dependencies than c	image captioning, which eliminates gradient decay—a characteristic found in even highly specialized RNN models, such as LSTM. RNN layer stacking. that offer extended sequence caption models improved performance. Rather than creating state-of-the-art model, our main attention has been on fixing the problems posed by RNN approaches in Image Captioning. As the field is investigated, methods for creating image captioning models are becoming more sophisticated. It enhances many practical uses, like as autonomous vehicles and image retrieval, and it can be applied for the general good.
16	Manish Raypurkar, et. al., 2021, [19]	CNN and RNN	A Recurrent Neural network (RNN)comes after a convolution neural network (CNN)in framework. The approach is able to produce the image caption that are typically grammatically correct and semantically meaningful by learning from image and caption pair.	Our model, which uses rapid text and CNN for multi-label classification, may be used to identify and extract objects from images and generate captions depending on the datasets that are supplied. Several methods, including convolution neural networks, long short-term memory, and recurrent neural networks, have been presented for the Image Caption Generator.

17	Prachi Waghmare, et. al., 2020 [20]	CNN and LSTM	Our goal in this work to understand a hybrid system that make use of an LSTM to efficiently arrange the relevant sentence utilizing the extracted or removed keyword and convolution neural network (CNN) to generate accurate descriptions of the photo.	In this paper we studied about the latest work done on captioning of image. Based on the dive's techniques adopted in each method, we categorized the captioning process in various groups. Numerous methods in each category are thoroughly elucidated, and their respective strengths and limitations are also addressed. We first discuss early captioning work of the image. Then, our main attention is focused on the overview of models. The challenges faced in image captioning and also a study of different articles on image captioning and at last we present a discussion on future research directions of automatic captioning process of images.
18	Priyanka Raut, et. al., 2021 [21]	Deep Learning, CNN and LSTM	Various solution has been employed to address the challenging task of generating concise sentences, known as caption, using neural networks. These methods still have a problem, though, such as incorrect caption producing captions just for viewed images, etc, the system model proposed in this paper, which consist of two stages and combines Deep neural network	The suggested Convolutional deep neural network extracts important features from the image and stores it in a feature vector. The feature vector is transmitted to the LSTM model for the generation of a sequential sentence, combing the extracted features and their relationship to form a caption. The proposed model generates accurate captions for the image and its error rate has been

			method (Convolution and Long Short-Term Memory) successfully produce more accurate caption. (3).	effectively reduced. The suggested system uses the LSTM algorithm to address the gradient vanishing problem inherent in the traditional RNN algorithm. The proposed model is trained with 6000 images from the Flicker8k dataset, which consists of images and their corresponding captions. 2000 photos from the Flicker8k dataset are used to test the system. The items and their relationships in the photos are being correctly identified by the system. This suggested model can be expanded in the future by adding additional descriptions, along with photos and captions, to train a system and increase the caption accuracy.
19	Jianhui Chen, et. al., 2014 [22]	CNN RNN and Sentence Generation	We examine the convolutional neural network (CNN), recurrent neural network (RNN), and phrase synthesis as the three main parts of the method. We discover that the VGGNet performs best based on the BLEU score after substituting three cutting-edge designs for the CNN portion. We also suggest using MATLAB and C++ with Caffe to construct a new recurrent layer that is a condensed version of the	We examined and adjusted the LRCN image captioning technique. We broke down the approach into its component parts-sentence generation, CNN, and RNN in order to fully comprehend it. We changed or swapped out each component to observe how it affected the outcome. The COCO caption corpus is used to assess the updated approach. The findings of the experiment indicate that: (1) VGGNet

			Gated Recurrent Units (GRU).	performs better in BLEU score measurement than AlexNet and GoogleNet; (2) the simplified GRU model achieves comparable results with more complex LSTM model; and (3) increasing the beam size generally raises the BLEU score but does not necessarily improve the quality of the description that is evaluated by humans.
20	M. Nivedita, et. al., 2029 [1]	Deep Learning, CNN and RNN	The primary task is accurately recognizing the objects present in the provided image. After recognizing all the objects and their attributes, the dense model undergoes training to identify the appropriate verbs or actions associated with recognize objects.	The suggested method is explained in full, along with how our image captioning model makes caption predictions for different types of affine modified images. Self-driving cars, image search, and visually challenged individuals can all benefit from it. For a subset of the affinely altered photos, the suggested approach yields appropriate captioning. All transformations except rotation were supported by our model. In the future, the work's accuracy will need to be increased in order to produce captions that are consistent for each modified image.

3. Problem Definition

To design a system capable to generating an accurate caption based on the Input Image using (CNN) Convolutional neural Network and (LSTM) Long Short-Term Memory algorithm. CNN is employed to extract features from the images. Convolutional Neural networks are specialized deep neural networks capable for processing the data with an input shape as like a 2D matrix. Images can be readily represented as a 2D matrix and CNN are highly effective for working with images. CNN are primarily utilized for

image classifications task, distinguishing and identifying objects within images such as determining whether an image as a bird, a plane or Superman, or other specified categories.
It scans images from left to right and top to bottom, extracting significant features and combining them to classify images. It can handle the images that undergo translation, rotation, scaling and changes in perspective.
Utilizing information from CNN, LSTM aid in the generation of image description. Long Short-Term memory (LSTM) is a subtype recurrent neural network of (RNN) specifically designed for effectively addressing sequence prediction challenge. Based on the previous text, we can predict what the next word will be. It has proven itself effective from the traditional RNN by overcoming the limitations of RNN which had short term memory. LSTM can retain pertinent information throughout the input processing, and though of forget gate, it filters out the non-relevant information.

3.1 Scope and Objectives of the System

3.1.1 Scope

- Utilizing CNN to extract features from the input image.
- LSTM utilizes information from CNN to assist to generating image descriptions.
- Image captioning is used for visually unpaired people to understand the surrounding situations.
- It helps people who visually unable to see by captioning image and convert it into another form such as audio.

3.1.2 Objectives

- To help visually unpaired people to understand the surrounding actions.
- To give meaningful caption of images.
- Easy to understand images by generating caption.

4. Methodology

The proposed system is divided into following Modules.

- The Input Image is first given as the input to the Image Based Module which uses CNN algorithm's Convolutional and Pooling layer, to generate a vector which is known as feature vector of the input image. After every Convolutional layer, a ReLu layer is used. And then Pooling layer is used to reduce the size of feature vector before passing it to the next model. The last layer of CNN, Fully Connected Network is excluded from this Model as we just need the feature vector. Convolutional and Pooling layer are used as Feature extractors whereas the Fully Connected Network is used as Classifier.
- Now, the output of previous model which is vector of features generated is given to next Module, Language Based Module where the encoded features vector is decoded into a natural language caption using LSTM, Long Short-Term Memory algorithm which is advanced version of Recurrent Neural Network has an advantage of storing long sequence of data. LSTM has a memory cell which can store the data for a longer period of time. The Sequence of sentence/caption has 2 special token which are start sequence and end sequence token so that the algorithm knows when to start the sequence and stop the sequence of sentence.
- Finally, a Caption is produced Color, actions and connection between these items are the main points of emphasis for the captioning model.

- Now make this caption communicate so that folks that are blind or visually challenged can recognize it.

4.1 Dataset

We have used Flicker8k Data set for model experiments. The model can be trained effectively using Flicker8k data set. The Flicker8k_dataset.zip file has:

- **Flickr8k_Dataset:** This folder has 8092 images, each image with different sizes, shapes and colors. From 8092 images, 6000 images are used for training, 1000 images are used for development and the rest 1092 images are used for the testing the proposed model. The size of this dataset is 1 GB.
- **Flickr8k_text:** The Flickr8k_text folder has a Flickr8k.token.txt file which has 5 captions per image for training the model is saved as a key-price pair, with the image's caption serving as the fee and the important thing being the image's precise identification. This document has a size of 2 MB.

4.2 Algorithm

Step 1: First, preprocess the input image.
Step 2: The enter photo's preprocessed pixel matrix is fed to a photograph-based version system, which creates a vector representation of the features the usage of CNN.
Step 3: Features of Output LSTM is used to decode vector right into a caption in herbal language. Transform this caption into speech in step four.

4.3 Modules of System

- **Image Preprocessing**

The photographs are not understood by the machine or gadget. The enter picture is first converted into a hard and fast-sized (224x224x3) pixel matrix, wherein every pixel's colour code is placed in its appropriate region. The noise in each and each photograph is then removed all through pre-processing. A threshold price is then defined to split the picture into foreground and heritage once it has been converted to grayscale. Every object in a photograph undergoes part detection. The output of the Image Pre-processing version and the input for the subsequent Module is the very last pixel matrix.

- **Image Based Model using CNN**

A redesigned CNN module that makes use of the convolutional and pooling layers to extract features. The matrix of pixels this is the output of the previous pre-processing module serves because the input for the image-primarily based module. This module creates a feature vector by using extracting capabilities from the picture pixel matrix. The convolutional layer is the initial CNN layer used in this module for feature extraction. Every convolutional layer is accompanied through a ReLU layer. The function vector is then contracted in length without sacrificing any of the photo's features by making use of a pooling layer. Features are retrieved from this module and saved in a characteristic vector. These functions consist of objects, verbs that describe the item's behavior, the object's coloration, and the maximum enormous dating among the object. The function vector, which is that this module's output, serves as the following modules enter. The vector's size is linearly transformed to the LSTM community's input size, that's utilized in the subsequent module

- **Language Based Module (LSTM)**

The Input to the Language Based Module is the linear feature vector for a given input image. The main aim of this module is to convert the encoded features into a simple language which can be understandable to the users using Long Short-Term Memory (LSTM). The module uses LSTM algorithm as it overcomes the variant gradient issue of the RNN algorithm and also can store a long sequence of data without forgetting the sequence of the data. The LSTM uses its memory cells to store the data. For training The LB i.e. Language Based Model, we first have to pre-define our label and target text. The Label stores the data in a sequence starting with a start token and the Target stores the sequence of data with an end token at the end of its token so that the algorithm understands when to stop.

- **Caption Generation**

The final module, caption technology, gets its input from the Language Based Module that got here before it. The purpose of this module is to generate a caption for an input photo in a linear style the use of the preceding module's commands. This module ends with the era of a caption in a format that is easily understood through people.

- **Text to Speech Conversion**

Lastly Caption based on Text converted into Speech so visually impaired people can hear and understand easily.

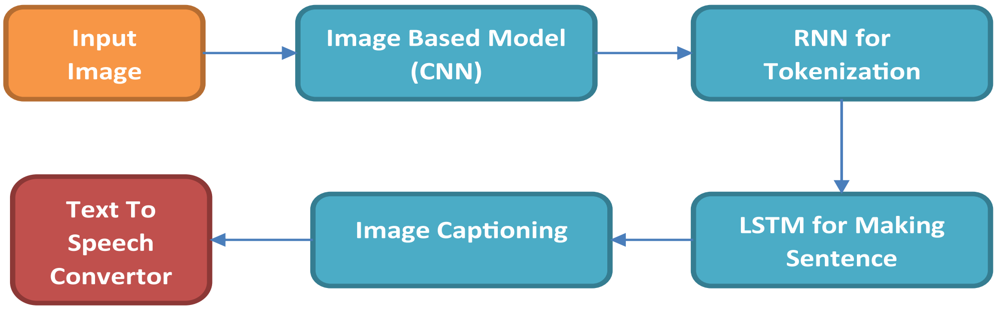

Figure 2: System Flow Diagram

5. Experimental Results

To verify the suggested approach, we can use $BLEU$ (Bilingual Evaluation Understudy). $BLEU$ metric is based on the precision measure. The precision of a sentence is calculated by dividing the total number of words in the candidate sentence by the number of consecutive words (n-grams).

More precisely, supposed to have a generated description G and a real description (reference) R, $BLEU$ score between G and R is computed as follows:

$$BLEU\ (N, G, R) = P(N, G, R) \times BP(G, R) \qquad (1)$$

where,

$P(N, G, R) = \left(\prod_{n=1}^{N} P_n\right)^{1/N}$ is the geometric mean of n-grams precision

$pn = {mn}/{ln}$ is the number of matched n-grams between G and R

ln is the total number of n-grams in G

$BP\ (G, R) = \min\left(1.0, \exp\left(1 - \left(\frac{len(R)}{len(G)}\right)\right)\right)$ is a brevity penalty in case the length of the generated sentence $len(G)$ is smaller than the one of reference $len(R)$.

In case there is no higher order n-gram precision (e.g. $n = 4$) in a sentence, the entire $BLEU$ score of the sentence is 0 independently from the quantity of the lower n-grams (n = 1,2,3) matching found in the sentence.

Table 1: Experimental Results

Input Image	Prediction	BLEU 1	BLEU 2
	child in red jacket is walking through the snow	0.444444	0.125000
	young boy in red shirt is jumping into the water	0.600000	0.111111
	dog is running through the water	0.833333	0.600000
	horse and rider are running on the track	0.375000	0.142857

6. Conclusion

Our model, that is primarily based on multi-label class using CNN and brief textual content to speech, is useful for extracting items from pictures and developing captions relying on the datasets which can be supplied. We have supplied several techniques for developing photo captions, along with recurrent neural networks, lengthy short-time period memory, and convolution neural networks.

Conflict of Interest: No potential conflict of interest was reported by the authors.

Acknowledgement

The authors are thankful to the Ipcowala Institute of Engineering and Technology (IIET), Dharmaj, Gujarat for the infrastructure and facilities.

References

1. Nivedita M., Phamila A.V.Y., “Image Captioning for Affine Transformed Images using Image Hashing”, International Journal of Engineering and Advanced Technology (IJEAT), 2019, 9 (1), 4736- 4741.
2. Reddy C., Reddy R., Sampath, Niteesh, Bhayana A., Jaisakthi S.M., Archit, “Deep Learning Based Image Caption Generator”, International Research Journal of Engineering and Technology (IRJET), 2021, 8 (12), 9-13.
3. Raut P., Deshmukh R.A., “An Advanced Image Captioning using combination of CNN and LSTM”, Turkish Journal of Computer and Mathematics Education, 2021, 12, 129-136.
4. Amritkar C., Jabade V., “Image Caption Generation using Deep Learning Technique”, Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, 1-4.
5. Lin J.Y., Chiang C.L., Wu M.J., Yao C.C., Chen M.C., "Smart Glasses Application System for Visually Impaired People Based on Deep Learning", Indo-Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN), Rajpura, India, 2020, 202-206.
6. Kim B., Shin S., Jung H., “Variational Autoencoder-Based Multiple Image Captioning Using a Caption Attention Map”, Applied Sciences. 2019, 9 (13):2699-2700.
7. Hoxha G., Melgani F., Demir B., "Retrieving Images with Generated Textual Descriptions”, IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 2019, 5812-5815.
8. Herdade S., Kappeler A., Soares J., Boakye K., “Image Captioning: Transforming Objects into Words.” Neural Information Processing Systems, 2019.
9. Liu S., Bai L., Hu Y., Wang H., “Image Captioning Based on Deep Neural Networks”, MATEC Web of Conferences, 2018, 232, 01052.
10. Wang C., Zhou Z., Xu L., “An Integrative Review of Image Captioning Research”, Journal of Physics: Conference Series, 2021, 1748 (4), 042060-042066.
11. Srinivasan L., Sreekanthan D., Amutha A.L., “Image Captioning-A Deep Learning Approach”, International Journal of Applied Engineering Research,2018, 13 (9), 7239-7242.
12. Hossain M.Z., Sohel F., Shiratuddin M.F., Laga H., “A Comprehensive Survey of Deep Learning for Image Captioning”, ACM Computing Surveys, 2017, 0(0), 0-7.
13. Wang H., Zhang Y., Yu X., "An Overview of Image Caption Generation Methods", Computational Intelligence and Neuroscience, 2020, 2020, 3062706.
14. Staniūtė R., Šešok D., “A Systematic Literature Review on Image Captioning”, Applied Sciences, 2019, 9 (10), 2024.
15. Haseeb S., Srushti G.M., Haripriya B., Prakash M., “Image Captioning using Deep Learning”, International Journal of Engineering Research & Technology (IJERT), 2019, 8 (5), 381-388.
16. Feng Y., Ma L., Liu W., Luo J., "Unsupervised Image Captioning," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 4120-4129.
17. Soh M., “Learning CNN-LSTM Architectures for Image Caption Generation”, 2016.
18. Rampal S., Gupta S., Verma S., Vishwakarma D.K., “Image Captioning using IndRNN”, International Journal of Advanced Science and Technology, 2020, 29 (08), 2211-2217.

19. Raypurkar M., Supe A., Bhumkar P., Borse P, Sayyad S., "Deep Learning Based Image Caption Generator", International Research Journal of Engineering and Technology (IRJET), 2021, 8 (3), 554-559.
20. Waghmare P., Shinde S., "Artificial Intelligence Based on Image Caption Generation", 2nd International Conference on Communication & Information Processing (ICCIP), 2020.
21. Raut P., Deshmukh R.A., "An Advanced Image Captioning using combination of CNN and LSTM", Turkish Journal of Computer and Mathematics Education, 2021, 12 (1S), 129-136.
22. Chen J, Dong W, Li M. "Image caption generator based on deep neural networks". https://www.math.ucla.edu/~minchen/doc/ImgCapGen.pdf.