Deep learning-based image captioning for visually impaired people

R. Kavitha1*, *S*. Shree Sandhya[2], *Praveena* Betes[2], *P*. Rajalakshmi[2], and *E*. Sarubala[2]

[1]Professor, CSE Department, Parisutham Institute of Technology and Science, India
[2]UG students, CSE Department, Parisutham Institute of Technology and Science, India

Abstract. Vision loss can affect people of all ages. Severe or complete vision loss may occur when the eye or brain parts that need to process images are damaged. In this paper, in order to facilitate the blind, deep learning algorithms are used to caption the image for the blind person in which the blind can know about the object, distance and position of object. Whenever an image is captured via the camera, the scenes are recognized and predicted by the machine. Afterthe prediction, it will be sent as an audio outputto the user. Thus, with the help of this paper an artificial vision to the blind, can be achieved and help them to gain confidence while travelling alone.

1 Introduction

Vision impairment may be due to sickness, an accident or a medical condition. This paper is aimed at providing assistance for the blind so that they can feel more confident, secured and independent. The paper involves developing a system that can automatically generate textual descriptions of images to allow blind individuals to better understand and interact with visual content. This system can be integrated into various platforms such as mobile applications, websites, and assistive devices, enabling visually impaired individuals to access information in a way that was previously not possible. To analyze images and to generate descriptive captions, machine vision, natural language generation and learning techniques are together utilized.

Convolutional neural networks (CNNs) for extracting picture features and recurrent neural networks (RNNs) for generating language are two deep learning techniques that can be used to complete the task. The two neural networks that are used in image captioning are CNN and RNN. CNN is a multi-layered neural network. It is designed to extractfeatures at each layer that are increasingly complex and to determine the output. CNNs are able to automatically extract meaningful visualfeatures from the input image, which can be used to generate more accurate and descriptive captions. The RNN is a multi-layered neural network that stores input in context nodes, enabling it to learn data sequences and produce a sequence as an output. RNNs are commonly used in image captioning as a decoder network to generate the corresponding caption from the visual features extracted by a CNN.

* Corresponding author : kavithha@gmail.com

Image caption generator is the technique of recognizing and understanding the context of an image and annotating it with relevant captions using deep learning. The goal of image captioning is to produce a coherent andsemantically meaningful description of animage that captures the main objects, actions, and attributes depicted in the visual content. There are numerous uses for image captioning in industries like assistive technology, image search, and social media. Datasets are used for generating the captions for images. The datasetused in this paper is MS-COCO (Microsoft Common Objects in Context).

1.1 Literature Survey

Literature survey in our paper represents the ground study of what we have done for the completion of the paper. It is a survey of previously published research papers on the topic of image captioning and deep learning-based research papers. [1] presents a text encoder, an image encoder, and a decoder to generate natural language descriptions of images. The model works in two stages, using a reinforcement learning algorithm to refine the initial caption generated in the first stage. [2] developed a summarization module that generates a summary of the remote sensing image, which is then used to guide the captioning process in the deep captioning module. The deep captioning module generates a natural language description of the remote sensing image based on the summary and the image itself. [3] It is an approach that involves training a neural network consisting of two main components: a visual feature extractor and a language generator. The visual feature extractor is typically a convolutional neural network (CNN) that is trained to extract features from the image. The language generator is a recurrent neural network (RNN) that generates the image caption word-by-word. [4] It is a method for generating more precise and detailed image captions by leveraging online positive recall and missing concepts mining. The approach uses a two-stage framework to generate captions that are more informative and accurate. [5] It is an approach that uses two-phase learning framework to generate captions. In the first phase, a visual saliency detector is trained to identify the salient regions of the image. In the second phase, a standard image captioning model is trained using the saliency maps generated by the saliency detector as additional input. [6] A cross-modal retrieval model is trained to learn a shared representation space for images and captions from both the source and target domains. A model adaptation technique is used to fine-tune the cross-modal retrieval model. [7] First extracts visual features from the input image using a CNN, and then encodes the visual features into a fixed-length vector using a recurrent neural network. [8] The encoded visual features are used as input to a context-aware policy network, which generates a sequence of words that describe the image. [9] It consists of two tasks: a source domain captioning task and a target domain captioning task. The source domain task is trained on a dataset of images from the same domain as the training data, while the target domain task is trained on a smaller dataset of images from the target domain. [10] Uses instance-level fine-grained feature representation and demonstrated its effectiveness through extensive experiments. [11] It consists of two main components: a generator network and a discriminator network. It combines generation- and retrieval-based methods using a dual generator generative adversarial network. [12] To improve the quality of the generated captions, a novel loss function is used that combines the attribute detection loss, the attribute prediction loss, and the captioning loss.

2 Materials and Methods

In Fig 1, Once the image is captured, it is firstdivided into ‘n’ pieces and then computes an image representation for each part. Using a compound coefficient, the EfficientNet-B3 isolates the image's features and uniformly scales the depth, width and resolution in all

three dimensions. Tokenization splits the input data into a sequence of meaningful parts. In tokenization, the image is split into patches andtext is split into tokens.

After tokenization, the extracted features will be fed for training with the Recurrent Neural Network (RNN) algorithm. A feedback connection is the vital characteristics of RNN. An RNN feedback loop has the ability to transfer the effects between the earlier portion and later portion of the sequences which is an essential capability of modeling the sequences. The MS-COCO dataset is used for understanding the visual scenes and generating captions.

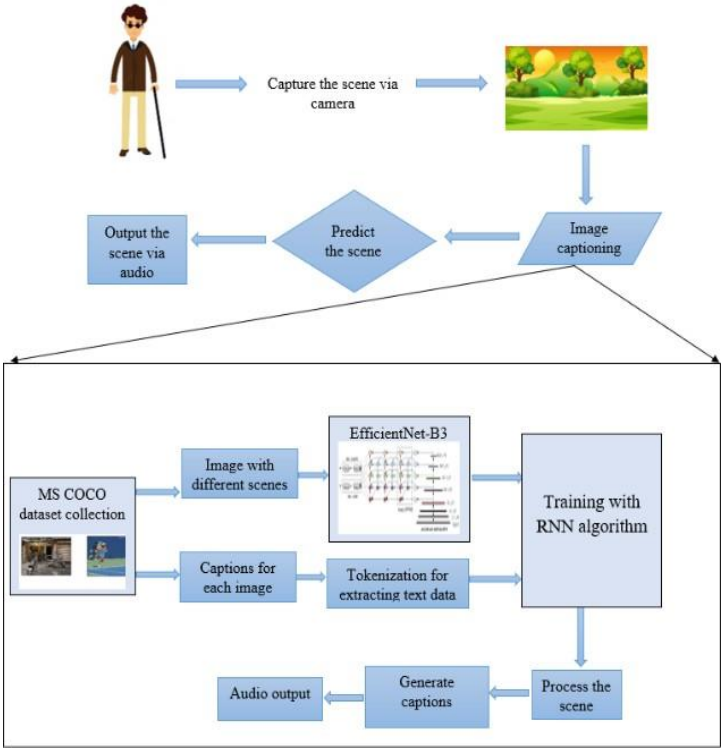

Fig 1. Architecture Diagram

2.1 Convolutional Neural Network (CNN)

The CNN is trained on a large dataset of images to learn a hierarchical representation of visual features. Once the CNN has extracted the features from the image, they are passed to a RNN, which is responsible for generating the captions. during training the input image is fed through the pre-trained CNN, and the output features from one or more of the intermediate layers are extracted. In Fig 2, the input layer, which receives the unprocessed picture data as input, is the initial layer of a CNN. The convolutional layer is the following layer, and it uses a series of filters to extract features from the input image. These filters are developed through training and are capable of identifying edges, corners, and other interesting aspects of the image. The fully connected layers employ the features that have been learned from the convolutional layers to assign the image to one of several categories or forecast a numerical value. CNNs are highly effective for image recognition tasks because they are able to automatically learn features from the input data.

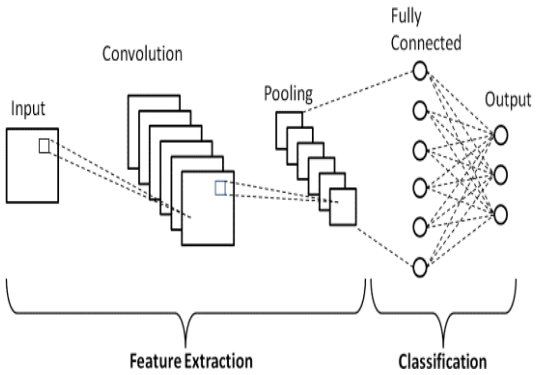

Fig 2. Convolutional Neural Network

2.2 Recurrent Neural Network (RNN)

RNN takes the features from the CNN as input and generates a sequence of words, one word at a time. RNNs are designed to handle sequential data, making them well-suited for generating sequences of words, such as captions. The RNN typically uses a typeof LSTM (Long Short-Term Memory) network, which is able to capture long-term dependencies in the sequence of visual features. In Fig 3, At each time step, the LSTM takes the output from the previous time step and combines it with the current input to generate a new output. This output is then passed through a fully connected layer to generate a probability distribution over the vocabulary of words. The final output of the RNN is a sequence of wordsthat describe the content of the image.

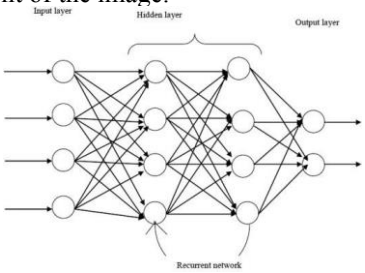

Fig 3. Recurrent Neural Network

2.3 Training Dataset

In machine learning model, a set of data from the entire data will be taken as training dataset. It consists of a collection of input data and their corresponding output values. MS COCO datasets are used in this paper. Exploring the MS COCO dataset, a sizable image dataset with 328,000 photos of common objects and people, is mostly done to comprehend the visual situations. The dataset is made up of the output captions for the input photos. The model may learn more precisely and generalise to new, untried data more effectively if the training data set is larger. The training dataset for supervised learning contains pairings of input and output data with the aim of teaching the machine learning model to translate the input data to the appropriate output data. In unsupervised learning, the training dataset simply contains input data, and the objective is to discover patterns or structure in the data. In Fig 4, the images in the MS COCO dataset cover a wide range of scenes and objects,

including people, animals, vehicles, and indoor and outdoor environments. This diversity makes the dataset well-suited for training machine learning models that can generate accurate and diverse captions for variety of images.

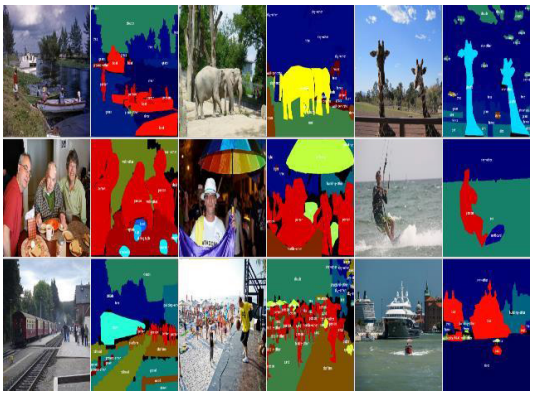

Fig 4. MS-Coco Dataset

2.4 Tokenization

Text data modeling starts with tokenization, Tokenization is the process of dividing a stream of textual data into tokens, which can be words, terms, sentences, symbols, or other significant objects. Unstructured data and text written in natural language are tokenized into informational units that can be regarded as separate elements. Tokens can either be words, character or subwords. In Fig 5, RNN uses the words that came before it to anticipate the subsequent words in a phrase. In order to achieve this, the tokenized word list in the caption of the image is transformed. Strings are turned to integers using the tokenization process. Create a dictionary that translates all distinct words to a numerical index by first going through all of the training captions. It will therefore have an integer value for each word it encounters.

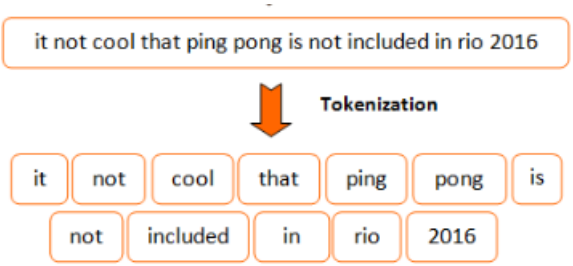

Fig 5. Tokenization

2.5 Caption Generation

Caption generation in image captioning refers to the process of generating a natural language description of an image. The goal is to train a machine learning model to automatically generate captions that accurately andsemantically describe the content of the input image. Generating accurate and semantically meaningful captions for images is a

challengingtask, and there are many techniques and approaches that have been developed to improve the performance of image captioning models.

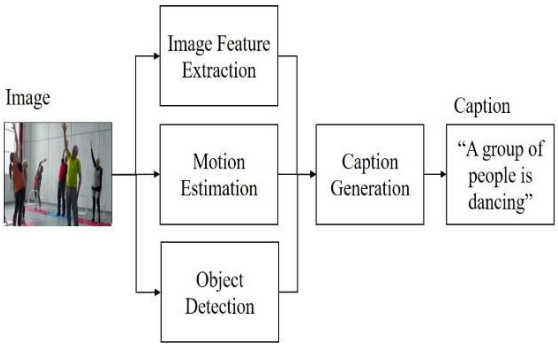

Fig 6. Caption Generation

These include using attention mechanisms to focus on specific regions of the image, incorporating external knowledge sources, andusing reinforcement learning to optimize the caption generation process. In Fig 6, The captions for the image are generated and produced as audio output that can help the blindpeople to identify the objects.

3 Results and Discussion

Once the image fed into the CNN, the image is first divided into 'n' pieces and then computes an image representation for each part. The EfficientNet-B3 separates the features of the image and scales the depth, width, and resolution in all three dimensions consistently. Tokenization splits the input data into a sequence of meaningful parts. In tokenization, the image issplit into patches and text is split into tokens. After tokenization, the extracted features willbe fed for training with the Recurrent Neural Network (RNN) algorithm. An RNN's main characteristic is its network of feedback links. This feedback loop gives the RNN the ability to simulate how the earlier portions of the sequence affect the later portions of the sequence, which is a crucial capability when modelling sequences. The MS-COCO dataset, which contains 328,000 pictures of people and objects from everyday life, is used to comprehend visual scenarios. The datasets will be collected from MS-COCO dataset and the datasets will be trained using advanced image captioning techniques implementing attention algorithm. Whenever an image is captured, the scenes are recognized and predicted by the machine. After training the model with algorithm, a live scene is captured via the camera. This captured scene will be recognized and the output model file will be generated. The major objects are also predicted and the distance is calculated from the camera. After the prediction, it is been sent as an audiooutput to the user.

4 Conclusion

In this paper, convolutional and recurrent neural networks, among other deep learning models, were investigated to provide captions for images. The use of pre-trained CNNs, such as the EfficientNet-B3 model, for feature extraction helped in capturing meaningful information from images, while the RNN generates sequential words to form coherent captions. The deep learning algorithm is the finest technique which ensures accuracy in the

achieved output. Overall, the study shows how deep learning methods may provide precise and insightful descriptions for images, which has benefits in areas like image retrieval and image indexing.

The datasets will be collected from MS-COCO dataset and the datasets will be trained using advanced image captioning techniques implementing attention algorithm. Whenever an image is captured via the camera, the scenes are recognized and predicted by the machine. After training the model with algorithm, a live scene is captured via the camera. This captured scene will be recognized and the output model file will be generated. The major objects are also predicted and the distance is calculated from the camera. As a result of the prediction an audio output has sent to the user.

References

1. Depeng Wang, Zhenzhen Hu, Yuanen Zhou, Richang Hong, *IEEE Transactions on Multimedia*, **23**, 3, pp. 779-789 (2022)
2. Sumbul G., Nayak S., & Demir B., *IEEE Transactions on Geoscience and Remote Sensing*, **59**, 8, pp. 6922-6934 (2021)
3. Yu N, Hu X, Song B, Yang J, Zhang J., *IEEE Transactions on Image Processing*, **28**, 6, pp. 2743-2754 (2019)
4. Zhang M, Yang Y, Zhang H, Ji Y, Shen H. T., Chua T., *IEEE Transactions on Image Processing*, **28**, 1, pp. 32-44 (2019)
5. Zhou L, Zhang Y, Jiang Y, Zhang T, Fan W., *IEEE Transactions on Image Processing*, **29**, pp. 694-709 (2020)
6. Zhao, W., Wu, X., Luo, J., *IEEE Transactions on Image Processing*, **30**, pp.1180-1192 (2021)
7. Maofu Liu, Huijun Hu, Lingjun Li, Yan Yu and Weili Guan, *IEEE Transactions on Cybernetics*, **52**, 2, pp. 1247- 1257 (2022)
8. Zheng-Jun Zha, Daqing Liu, Hanwang Zhang, Yongdong Zhang, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**, 2, pp. 710- 722 (2022)
9. Min Yang, Wei Zhao, Wei Xu, Yabing Feng,Zhou Zhao, Xiaojun Chen, Kai Lei, *IEEE Transactions on Multimedia*, **21**, 4, pp. 1047-1061 (2019)
10. Qingbao Huang, Yu Liang, Jielong Wei, YiCai, Hanyu Liang, Ho-fung Leung, Qing Li, *IEEE Transactions on Multimedia,* **24**, pp. 2004-2017 (2022)
11. Min Yang, Junhao Liu, Ying Shen, Zhou Zhao, Xiaojun Chen, Qingyao Wu, ChengmingLi, *IEEE Transactions on Image Processing*, **29**, pp. 9627-9640 (2020).
12. Yiqing Huang, Jiansheng Chen, Wanli Ouyang, Weitao Wan, Youze Xu, *IEEE Transactions on Image Processing*, **29**, pp. 4013-4026 (2020)