International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Image Caption for the Blind Using Deep Learning Algorithm

S. G. Prithivi Raj[1], Dr. V. Vaidehi[2]

PG Student[1], Professor[2]
Department of Computer Applications, Dr. M.G.R. Educational and Research Institute, Chennai, India.
E-mail: prithrocknroll@gmail.com[1], vaidehi.mca@drmgrdu.ac.in[2]

ABSTRACT-

Globally, 1 billion people have a vision impairment that could have been prevented or has yet to be addressed. This 1 billion people include those with moderate or severe distance vision impairment or blindness due to unaddressed refractive error (123.7 million), cataract (65.2 million), glaucoma (6.9 million), corneal opacities (4.2 million), diabetic retinopathy (3 million), and trachoma (2 million), as well as near vision impairment caused by unaddressed presbyopia (826 million). In terms of regional differences, the prevalence of distance vision impairment in low- and middle-income regions is estimated to be four times higher than in high-income regions. With regards to near vision, rates of unaddressed near vision impairment are estimated to be greater than 80% in western, eastern and central sub-Saharan Africa, while comparative rates in high-income regions of North America, Australasia, Western Europe, and of Asia-Pacific are reported to be lower than 10%. Population growth and ageing are expected to increase the risk that more people acquire vision impairment. In this project in order to facilitate the blind we will be using deep learning algorithm such as efficientnet B3 to caption the image for the blind person in which the blind can know about the object detection, distance and position of object. This is been achieved by using advanced image captioning techniques implementing efficientnet B3 algorithms and tokenization methods where the scenes with different captions are learned by the machine. Whenever an image is captured via the camera are been recognized and predicted by the machine. The major objects are also predicted and the distances calculated from the camera. After the prediction, it is been sent as an audio output to the user which can help them identify the distance and position of object. Thus, with the help of this project we provide an artificial vision to the blind, which can help them gain confidence while travelling alone.

Keywords- Recurrent Neural Networks(RNN), ImageNet Large Scale Visual Recognition Challenge(ILSVRC**),** Supervisory Control And Data Acquisition(SCADA)**,** Machine-to-Machine(M2M**),** Convolutional Neural Networks(CNN**)**

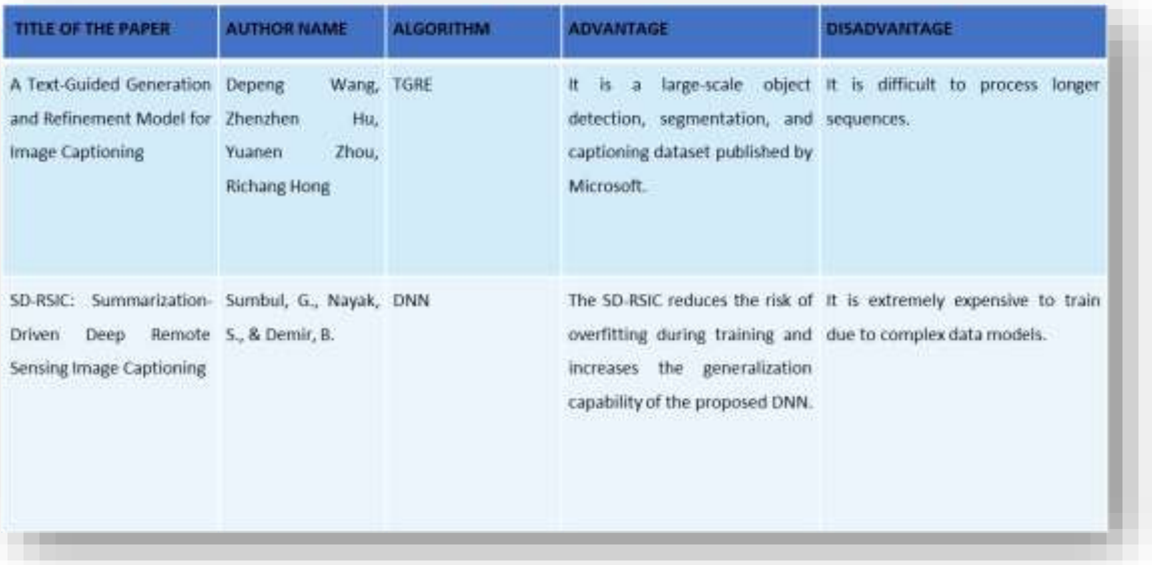
I. Introduction-

The human eye is like a camera that collects, focuses, and transmits light through a lens to create an image of its surroundings. In a camera, the image is created on film or an image sensor. In the eye, the image is created on the retina, a thin layer of light-sensitive tissue at the back of the eye

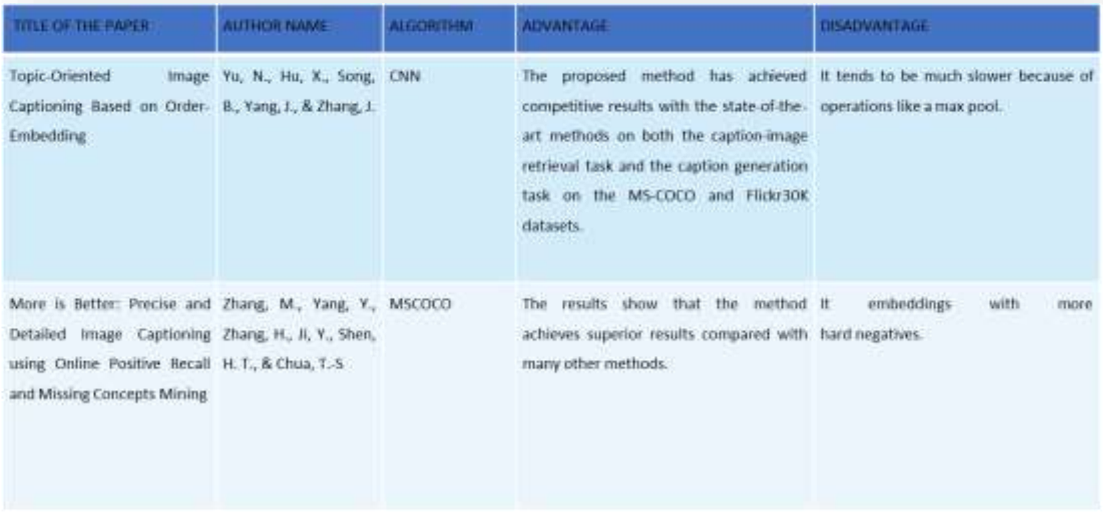
Like a camera, the human eye controls the amount of light that enters the eye. The iris (the colored circular part of the eye) controls the amount of light passing through the pupil. It closes up the pupil in bright light and opens it wider in dim light. The cornea is the transparent, protective surface of the eye. It helps focus light, as does the lens, which sits just behind the iris. When light enters the eye, the retina changes the light into nerve signals. The retina then sends these signals along the optic nerve (a cable of more than 1,000,000 nerve fibers) to the brain. Without a retina or optic nerve, the eye can't communicate with the brain, making vision impossible.

Many people have some type of visual problem at some point in their lives. Some can no longer see objects far away. Others have problems reading small print. These types of conditions are often easily treated with eyeglasses or contact lenses. But when one or more parts of the eye or brain that are needed to process images become diseased or damaged, severe or total loss of vision can occur. In these cases, vision can't be fully restored with medical treatment, surgery, or corrective lenses like glasses or contacts.

II. Literature Survey-



TITLE OF THE PAPER	AUTHOR NAME	ALGORITHM	ADVANTAGE	DISADVANTAGE
A Text-Guided Generation and Refinement Model for Image Captioning	Depeng Wang, Zhenzhen Hu, Yuanen Zhou, Richang Hong	TGRE	It is a large-scale object detection, segmentation, and captioning dataset published by Microsoft.	It is difficult to process longer sequences.
SD-RSIC: Summarization-Driven Deep Remote Sensing Image Captioning	Sumbul, G., Nayak, S., & Demir, B.	DNN	The SD-RSIC reduces the risk of overfitting during training and increases the generalization capability of the proposed DNN.	It is extremely expensive to train due to complex data models.



TITLE OF THE PAPER	AUTHOR NAME	ALGORITHM	ADVANTAGE	DISADVANTAGE
Topic-Oriented Image Captioning Based on Order-Embedding	Yu, N., Hu, X., Song, B., Yang, J., & Zhang, J.	CNN	The proposed method has achieved competitive results with the state-of-the-art methods on both the caption-image retrieval task and the caption generation task on the MS-COCO and Flickr30K datasets.	It tends to be much slower because of operations like a max pool.
More is Better: Precise and Detailed Image Captioning using Online Positive Recall and Missing Concepts Mining	Zhang, M., Yang, Y., Zhang, H., Ji, Y., Shen, H. T., & Chua, T.-S	MSCOCO	The results show that the method achieves superior results compared with many other methods.	It embeddings with more hard negatives.

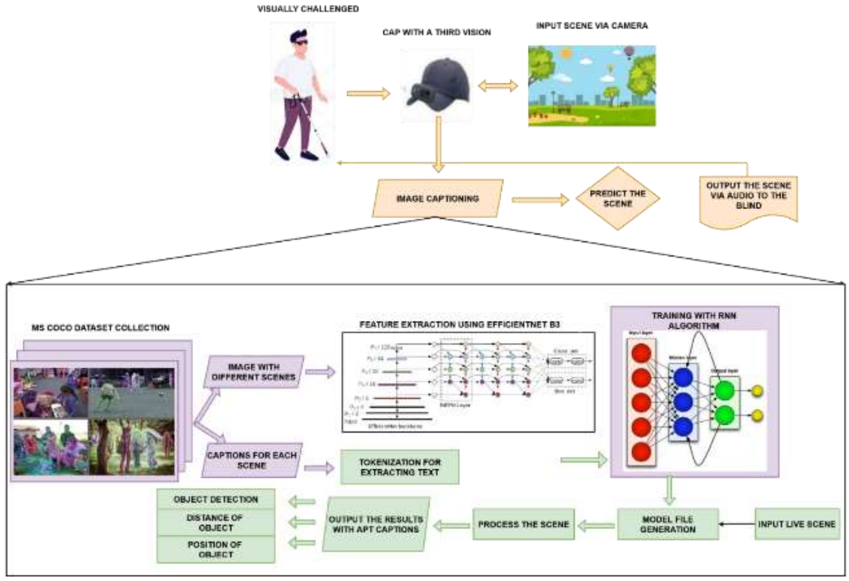
III. Proposed System-

In this project in order to facilitate the blind we will be using deep learning algorithm such as efficientnet B3 to caption the image for the blind person in which the blind can know about the object detection, distance and position of object. This is been achieved by using advanced image captioning techniques implementing efficientnet B3 algorithms and tokenization methods where the scenes with different captions are learned by the machine. Whenever an image is captured via the camera are been recognized and predicted by the processor. The major objects are also predicted and the distances calculated from the camera. After the prediction, it is been sent as an audio output to the user which can help them identify the distance and position of object. Thus, with the help of this project we provide an artificial vision to the blind, which can help them gain confidence while travelling alone.

IV Advantages of Proposed System-

Cheap and effective solution for the blind to become aware of the surroundings, Effective scene prediction model is developed, Provides an artificial vision to the visually challenged people.

V. Architecture diagram-


VISUALLY CHALLENGED
CAP WITH A THIRD VISION
INPUT SCENE VIA CAMERA
IMAGE CAPTIONING
PREDICT THE SCENE
OUTPUT THE SCENE VIA AUDIO TO THE BLIND
MS COCO DATASET COLLECTION
IMAGE WITH DIFFERENT SCENES
CAPTIONS FOR EACH SCENE
FEATURE EXTRACTION USING EFFICIENTNET B3
TOKENIZATION FOR EXTRACTING TEXT
TRAINING WITH RNN ALGORITHM
MODEL FILE GENERATION
INPUT LIVE SCENE
PROCESS THE SCENE
OUTPUT THE RESULTS WITH APT CAPTIONS
OBJECT DETECTION
DISTANCE OF OBJECT
POSITION OF OBJECT

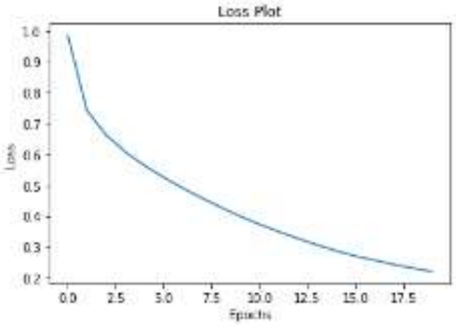
VI. Description-

In this project in order to facilitate the blind we have develop an artificial vision system with the help of deep learning techniques. Initially COCO dataset is collected to train the model, before training the model pre-processing of dataset is done so that the dataset can be directly applied to the deep Learning algorithms for further processes. After pre-processing feature extraction in done using efficientnet algorithm. Once the feature extraction is completed the model is trained using RNN algorithm. Recurrent neural networks (RNN) are the state-of-the-art algorithm for sequential data and are used by Apple's Siri and and Google's voice search. It is the first algorithm that remembers its input, due to an internal memory, which makes it perfectly suited for machine learning problems that involve sequential data. It is one of the algorithms behind the scenes of the amazing achievements seen in deep learning over the past few years. The major objects are also predicted and the distances calculated from the camera. This is achieved by using advanced image captioning techniques learned by the model. Whenever an image is captured via camera, the scenes are recognized and predicted by the machine. After the prediction, it is been sent as an audio output to the user which can help them identify the distance and position of object.

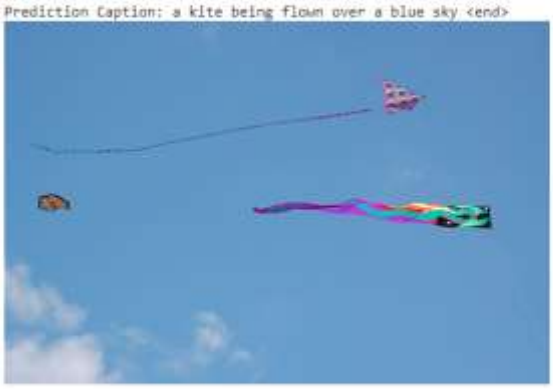
VII. Result-

To begin with, testing of the trained model, we can split our project into modules of implementation that is done. Dataset collection involves the process of collecting image caption dataset. Then these datasets are pre-processed from convert the images into required size format so that it can be made ready for training with the model. The below figure shows the simple pre-processing techniques used for image resizing.

Epochs vs Loss graph-



The below figure shows the validating of image caption-



```
Prediction Caption: a kite being flown over a blue sky <end>
```

VIII. Conclusion-

The project has been successfully implemented to provide a solution for the blind person in which the blind can have a device with them and know about the object detection, distance and position of object using the deep learning algorithm. The algorithm such as EfficientNetB3 to caption the image for blind and also identify the distance and position of object. The deep learning algorithm is the finest technique which ensures accuracy in the achieved output and algorithm has a highest quality guesswork. Thus, in this project in order to facilitate the blind we develop an artificial vision system which can help them gain confidence while travelling alone.

[XI]Future Work-

In the coming future, we will review the scope of this of the project in the medical field, and try to enhance this technique in other fields. And there are more chance to develop or convert this project in many ways. Thus, this project has an efficient scope in coming future to guide the blind people about the surrounding environment and also cheap and effective solution for the blind to become aware of the surroundings.

IX. References-

[1] Depeng Wang, Zhenzhen Hu, Yuanen Zhou, Richang Hong, (2022). A Text-Guided Generation and Refinement Model for Image Captioning. IEEE TRANSACTIONS ON MULTIMEDIA, DOI 10.1109/TMM.2022.3154149

[2] Huang, Q., Liang, Y., Wei, J., Yi, C., Liang, H., Leung, H., & Li, Q. (2021). Image Difference Captioning with Instance-Level Fine-Grained Feature Representation. IEEE Transactions on Multimedia, 1–1. doi:10.1109/tmm.2021.3074803

[3] Huang, Y., Chen, J., Ouyang, W., Wan, W., & Xue, Y. (2020). Image Captioning With End-to-End Attribute Detection and Subsequent Attributes Prediction. IEEE Transactions on Image Processing, 29, 4013–4026. doi:10.1109/tip.2020.2969330

[4] Liu, M., Hu, H., Li, L., Yu, Y., & Guan, W. (2020). Chinese Image Caption Generation via Visual Attention and Topic Modeling. IEEE Transactions on Cybernetics, 1–11. doi:10.1109/tcyb.2020.2997034

[5] Sumbul, G., Nayak, S., & Demir, B. (2021). SD-RSIC: Summarization-Driven Deep Remote Sensing Image Captioning. IEEE Transactions on Geoscience and Remote Sensing, 59(8), 6922–6934. doi:10.1109/tgrs.2020.3031111

[6] Wu, J., Chen, T., Wu, H., Yang, Z., Luo, G., & Lin, L. (2020). Fine-Grained Image Captioning with Global-Local Discriminative Objective. IEEE Transactions on Multimedia, 1–1. doi:10.1109/tmm.2020.3011317

[7] Xian, Y., & Tian, Y. (2019). Self-Guiding Multimodal LSTM -when we do not have a perfect training dataset for image captioning. IEEE Transactions on Image Processing, 1–1. doi:10.1109/tip.2019.2917229

[8] Yang, M., Liu, J., Shen, Y., Zhao, Z., Chen, X., Wu, Q., & Li, C. (2020). An Ensemble of Generation- and Retrieval-Based Image Captioning With Dual Generator Generative Adversarial Network. IEEE Transactions on Image Processing, 29, 9627–9640. doi:10.1109/tip.2020.3028651

[9] Yang, M., Zhao, W., Xu, W., Feng, Y., Zhao, Z., Chen, X., & Lei, K. (2018). Multitask Learning for Cross-Domain Image Captioning. IEEE Transactions on Multimedia, 1–1. doi:10.1109/tmm.2018.2869276

[10] Yu, J., Li, J., Yu, Z., & Huang, Q. (2019). Multimodal Transformer with Multi-View Visual Representation for Image Captioning. IEEE Transactions on Circuits and Systems for Video Technology, 1–1. doi:10.1109/tcsvt.2019.2947482