***CAPTION BOT FOR ASSISTIVE VISION***

***Authors:* Srinivas V 1BY17CS170, Rashid Pasha 1BY17CS175, Vimal Kumar 1BY17CS187,Shawez Khan 1BY17CS155**

**Abstract**

An In navigating and understanding an outdoor environment, our world often requires the ability to see. People with visual impairments are, therefore, faced with significant challenges in exploring these environments. Deep learning has the potential to alleviate part of the frustrations they face. In this thesis, we assess the effectiveness of using deep learning to assist people with visual impairments.

We are very interested in how machines can automatically describe the content of images using human language. In order to gain a deeper insight of this computer vision topic, we decided to implement current state-of-the-art image caption generator Show, attend and tell: Neural image caption generator with visual attention. Our neural network based image caption generator is implemented in Python powered by Pytorch machine learning library. We have identified five major components in our pipeline: (R1) data prepossessing; (R2) Convolution

Neural Network (CNN) as an encoder; (R3) attention mechanism; (R4) Recurrent

Neural Network (RNN) as a decoder; (R5) Sentence Generation and evaluation.

**Introduction**

The problem of generating natural language descriptions of an image to describe the visual content has received much interest in the fields of computer vision and natural language processing, driven by applications such as image indexing or retrieval, virtual assistants, image understanding and support of the visually impaired people. Although the visually impaired people use other senses such as hearing and touch to recognize the events and objects around them, the life quality of those people can be dramatically lower than standard level. For this reason, studies such as “guide dog” [1], “smart glasses” [2] and “image captioning” [3] are reported in order to improve the life quality of visually impaired. In this study, a new captioning approach is reported to describe visual content of an image which can be integrated to hardware platforms such as smartphone and smart glass in order to make their life not simply accessible but a socially meaningful and enjoyable experience. To generate a natural language description of an image, sophisticated algorithms are required that goes beyond image classification and object detection which attracts the interest of two major areas of artificial intelligence (AI): computer vision and natural language processing (NLP) [4]. NLP is defined as the automatic exchange of natural language such as general speech and text by software [5], and is a collective term referring to the automatic computational processing of human languages. This term includes both algorithms that take human-generated text as input and algorithms that produce natural-looking text as output [6]. Earlier techniques were designed to use statistical methods in NLP studies. However, theoretical and algorithmic advances together with the increasing capability in computer processing have led to the emergence of more sophisticated techniques like neural networks replaced by statistical methods [7]. Neural networks consist of extremely complex structures, however, deep learning methods provide an effective solution for the processing of data in these structures.

**Proposed Method**

The proposed system aims at performing real time object detection for visually impaired people. This can be achieved by implementing machine learning and image classification techniques. Artificial Neural Networks, and Convolution Neural Networks in specific, are one of the most accurate methods in order to achieve the desired result.

The basic ideology behind the proposed system/model is to develop a object detection model that is trained using a publicly available dataset. The model is trained using the various images present in the dataset, for which convolution neural networks is used. The trained model is then tested using the test dataset in order to test the working and prediction accuracy of the model. It consists of more than thirteen thousand images that have been captured in various scenarios, thus helpful in order to build a wholesome and accurate object detection model.

* Step 1: Object detection In this step, the objects in the input image are detected using R-CNN region proposal approach.
* Step 2: Feature Extraction In this step, the features in the image are extracted using principal component analysis using NumPy. CNN is used for scene classification and RNN is used for detecting objects and human attributes.
* Step 3: Creating attributes In this step, the features extracted by the neural networks were used to define the attributes with its label strings.
* Step 4: Encoder and Decoder In this step, the label strings were subjected to an encoder RNN for encoding the label strings to a proper format, and the resultant variable length string is subjected to a fixed length decoder for converting to a fixed length descriptive sentence.

**Steps of CNN**

Step 1: Convolution layer The Kernel

* + - * The role of the ConvNet is to reduce the images into a form which is easier to process, without losing features which are critical for getting a good prediction
      * if input image size is 5\*5 dimension and Kernel/filter dimension is 3\*3

then the output i.e. convolved feature is 3\*3 dimension(so we have reduced the image dimension)

Step 2: Pooling layer.

* + - * Similar to the Convolutional Layer, the Pooling layer is responsible for reducing the spatial size of the Convolved Feature. This is to decrease the computational power required to process the data through dimensionality reduction. Furthermore, it is useful for extracting dominant features which are rotational and positional invariant, thus maintaining the process of effectively training of the model.

Step 3: Flattening.

* + - * Now that we have converted our input image into a suitable form for our Multi-Level Perceptron, we shall flatten the image into a column vector. The flattened output is fed to a feed-forward neural network(ANN) and back propagation applied to every iteration of training.

Step 4: Full Connection.

* + - * Adding a Fully-Connected layer is a (usually) cheap way of learning non-linear combinations of the high-level features as represented by the output of the convolutional layer. The Fully-Connected layer is learning a possibly non-linear function in that space.
      * From the training dataset(Labelled images) our CNN will identify and learn the features of images and when a new image is given to our CNN will classify into a right category.
      * Once CNN classifies an image then we will convert the text (category) into voice so that the person can easily come to know what is in front of him/her.

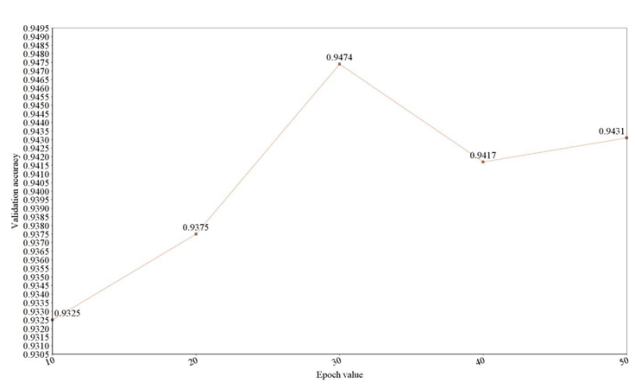
**Conclusion**

After going through the entire process of researching, implementing and testing the developed model, it can be concluded that Convolution Neural Network Technique proves to be a very efficient method for the detection and classification of objects in real time.

One of the most important findings was during the training process. It was found that model training failed if the training process was interrupted in between. Due to the

nature of the training process and the characteristics of the algorithm being used, it is necessary that the training process remains uninterrupted in order to make sure that model training completes successfully.

Another important finding was encountered while building the model. As mentioned in Table 5.3.1, the model was tested with different epoch values in order to monitor its performance at various stages. One epoch is when an entire dataset is passed forward and backward through the neural network only once. A training process may consist of more than one epoch. As the number of epochs increases, a greater number of times the weights are changed in the neural network.



**References**

1. Abedin, Tingke Shen, Amlan Kar, Sanja Fidle. " Learning to Caption Images Through a Lifetime by Asking Questions." *2019 19th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2019.
2. Behrendt, [Hongje Seong](https://ieeexplore.ieee.org/author/37087015739); [Junhyuk Hyun](https://ieeexplore.ieee.org/author/37085631260); [Euntai Kim](https://ieeexplore.ieee.org/author/37277101900) "A deep learning approach to object: Detection, tracking, and classification." *2019 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.
3. Qiuyuan Huang ,Pengchuan Zhang Dapeng ,WuLei Zhang et al. " Turbo Learning for CaptionBot and DrawingBot *2020 International Conference on Control, Automation and Diagnosis (ICCAD)*. IEEE, 2020.
4. Jung, Seokwoo, et al. "Real-time Traffic Sign Recognition system with deep convolutional neural network." *2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*. IEEE, 2016.

Li, Xi, et al. "Traffic light recognition for complex scene with fusion detections." *IEEE Transactions on Intelligent Transportation Systems* 19.1 (2018): 199-208