Accuracy in Object Detection Using Deep Learning Method

Divya Sirala and Kuldeep Singh Nagla

October 26, 2022

Accuracy in Object Detection by using Deep Learning method

Divya Sirala[1*], Kuldeep Singh Nagla [2]

Dr. B. R. Ambedkar National Institute of Technology, Jalandhar

[1] divyas.ic.20 @nitj.ac.in , [2] naglaks@nitj.ac.in

Abstract

Accuracy and Precision measurement of an object in complex environment for mobile robot applications in an important task. In complex environment it is difficult to identify similar objects, such as chair, table, furniture and other objects available in the indoor environment. Different methods are available to identify such objects, but high accuracy and precision is not easily achievable. This paper presents MobileNet with Vision Transformer for identification of objects in complex environment. Several experiments have been conducted for indoor environment where 94.3% accuracy is achieved for indoor environment with the implementation of MobileNet with Vision Transformer model.

Keywords: MobileNet, Vision Transformer, Mobile Robot, Accuracy and Precision measurement.

I. Introduction

For Mobile Robot, the process of object detection is an important task as it is what helps the autonomous robot to detect hurdles and react to its environment so as to navigate without the help or involvement of humans. Object Detection and Recognition method using deep learning is important as to verify new theories and methods for solving the various detection problem of general complex environment because of the accuracy and speed of the methods.

There are many other approaches like laser based detection, camera based detection, deep learning based detection, etc that might be used for object detection for mobile robots. However, Deep Learning has recently become quite popular because of its superior accuracy when taught on massive amounts of data.

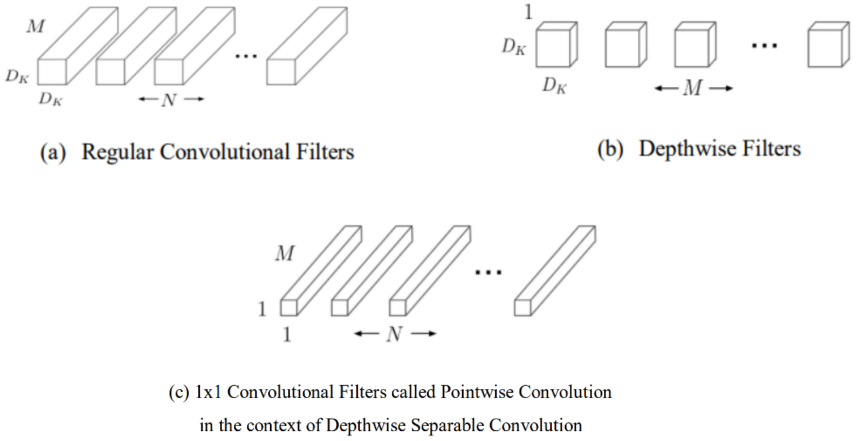
Although there are many Machine Learning and Deep learning based algorithms used for Object Detection and Recognition, such as Convolution Neural Networks(CNNs), Support Vector Machine(SVM), Regional Convolutional Neural Networks(R-CNNS), You Only look Once(YOLO) model etc., it becomes crucial to select the appropriate algorithm for Mobile Robot Object Detection which should solve the problem regarding speed, accuracy response time and input data used to train and test the model.

With the use of a custom dataset, the precision to the type of objects that need to be detected is increased. MobileNet model is used for object detection with custom dataset. Creating a light deep learning model with depthwise convolutions is the foundation of MobileNets, which are faster compared to other CNN architectures. Due to its smaller size and low latency, MobileNet outperforms other Deep Learning methods for object detection.

This proposed method describes an efficient network architecture using MobileNet with an additional Vision Transformer layer to be trained on a Custom Dataset of indoor environment for a Mobile Robot. The vision transformer divides an image into fixed-size patches, which it then accurately integrates by performing positional embedding's on each patch as an argument to the transformer encoder. By means of high accuracy and efficient computing, ViT models surpass CNNs by roughly four times [1]. With demonstration of experiments to detect certain objects present in the indoor environment, the factors determining the efficacy of the suggested model in comparison to a traditional model of MobileNet method for object detection are Accuracy, Precision, f1-Score, Confusion Matrix and Losses.

II. MobileNet

The depthwise separable convolutions, a type of factorised convolutions that factorise a conventional convolution into a depthwise convolution as well as a 1x1 convolutional known as a point - wise convolution, are the foundation of the MobileNet model. Using depthwise convolution, MobileNets applies a single filter to each input channel. The outputs of the depthwise convolution are combined using a 1x1 convolution after the pointwise convolution. In one step, a conventional convolution filters mixes inputs to invent a fresh set of outputs. This is divided into two layers by the depthwise separable convolution: a layer for combining and a layer for filtering. The computations and model size are significantly decreased as a result of this factorization. Fig 1.1 [2] illustrates the factorization of a conventional convolution into a depthwise and a 1x1 pointwise convolution


M
D_K
D_K
$\leftarrow N \rightarrow$
1
D_K
D_K
$\leftarrow M \rightarrow$
M
1
1
$\leftarrow N \rightarrow$

(a) Regular Convolutional Filters

(b) Depthwise Filters

(c) 1x1 Convolutional Filters called Pointwise Convolution in the context of Depthwise Separable Convolution

Fig 1.1 The standard convolutional filters in (a) are replaced by two layers: depthwise convolution in (b) and pointwise convolution in (c) to build a depthwise separable filter. [2]

III. Vision Transformer

Recently, there has been an increased interest in Vision Transformers (ViTs) in computer vision research. The introduction of self-attention for visuals is a key component of ViT's ideation. As a result, self-attention in spatial aspects makes sense. The ViT is a visual representation of the architecture of a transformer that was originally developed for text-based operations. The ViT model predicts the label for use with a classifier head and represents an input picture as a sequence of image patches, much as the order of word representations used when employing transformers to text [1]. ViT exhibits exceptional performance when given the necessary duties to execute and is well-versed in large data. With very few 4x computational needs, it can render contemporary CNN uses. These transformers have a good success rate for NLP models and are now utilized on photos for image recognition applications. ViT divides the pictures into visual tokens while CNN uses the pixels as arrays.

The visual transformer separates the pictures into fixed-size patches, properly integrate each one, and then provides the transformer encoder with the spatial embedding as input. Additionally, after each block, residual connections are offered since they enable components to move directly across the network without having to go through non-linear activations. When classifying images, the classification head is implemented by the MLP layer. It uses a single linear layer for fine-tuning and a single hidden layer for pre-training. The vision transformer paradigm in computer vision makes use of multi-head self-attention without the necessity for image-specific biases. The transformer encoder processes the positional embedding patches that the model creates from the pictures. It does this in order to comprehend the local and global aspects of the picture. Despite the need for image-specific biases, the vision transformer paradigm in computer vision leverages multi-head self-attention. The transformers encoder processes the positional integrating patches that the model creates from the pictures. It accomplishes this in order to comprehend the features both locally and globally present in the image. Lastly the ViT requires less training time and has a better accuracy rate on a big dataset.

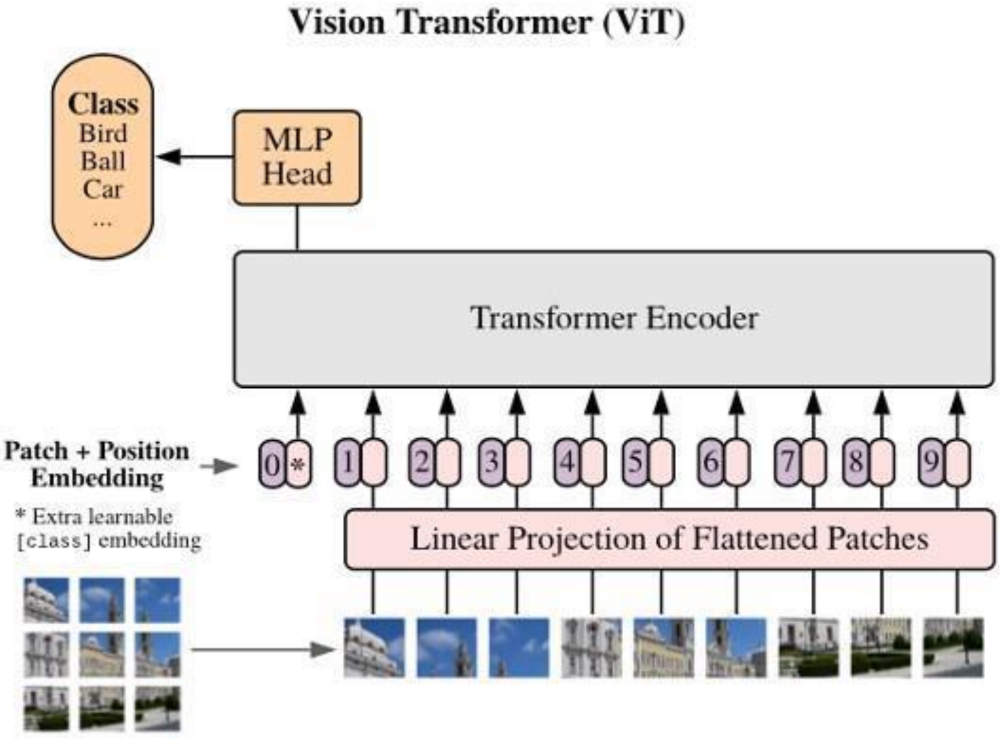

Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Fig 1.2 Vision Transformer Architecture [3]

I. Experimental Setup

Dataset Collection and Preparation – The Dataset was created for five different objects: Almirah, Table, Chair, Bag and Bottle in an indoor environment for Mobile Robot. This dataset was used to train and test both the neural networks with eighty percent of images used for training the model and twenty percent of the images were used for validation, i.e. testing the trained models. To generate a large variety of data, different backgrounds, object texture and object rotation and camera locations were used. Most of the images were downloaded from google in bulk, later using 'LabelImg' the annotations of all the images with bounding boxes were done.

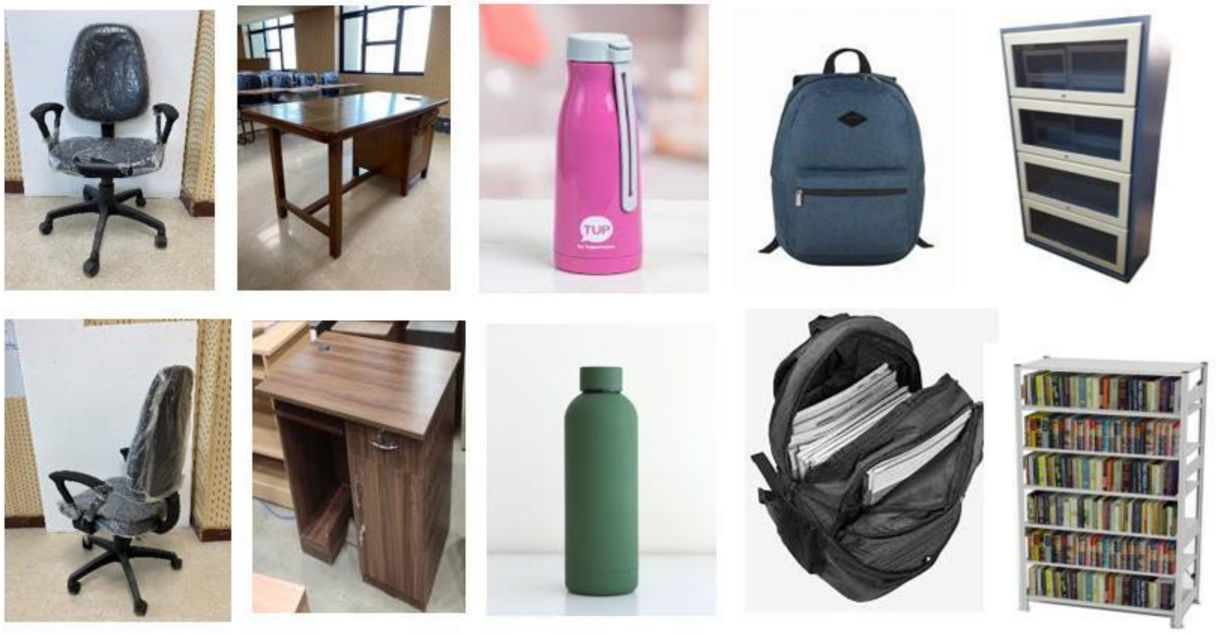


Fig 1. 3 Images used for Validation

Object detection models including convolutions and transformers can do better with a few data augmentations. Although the data provided may not contain sufficient images to train the model, modifying the already existing images in the dataset using data augmentation techniques can help the model learn more rigorously and reduce bias. Here two basic augmentation techniques i.e. image rotation and image protection are applied to fifty percentage of the total input data. The augmentation methods mentioned above help the model to integrate the problem much better and can increase performance.

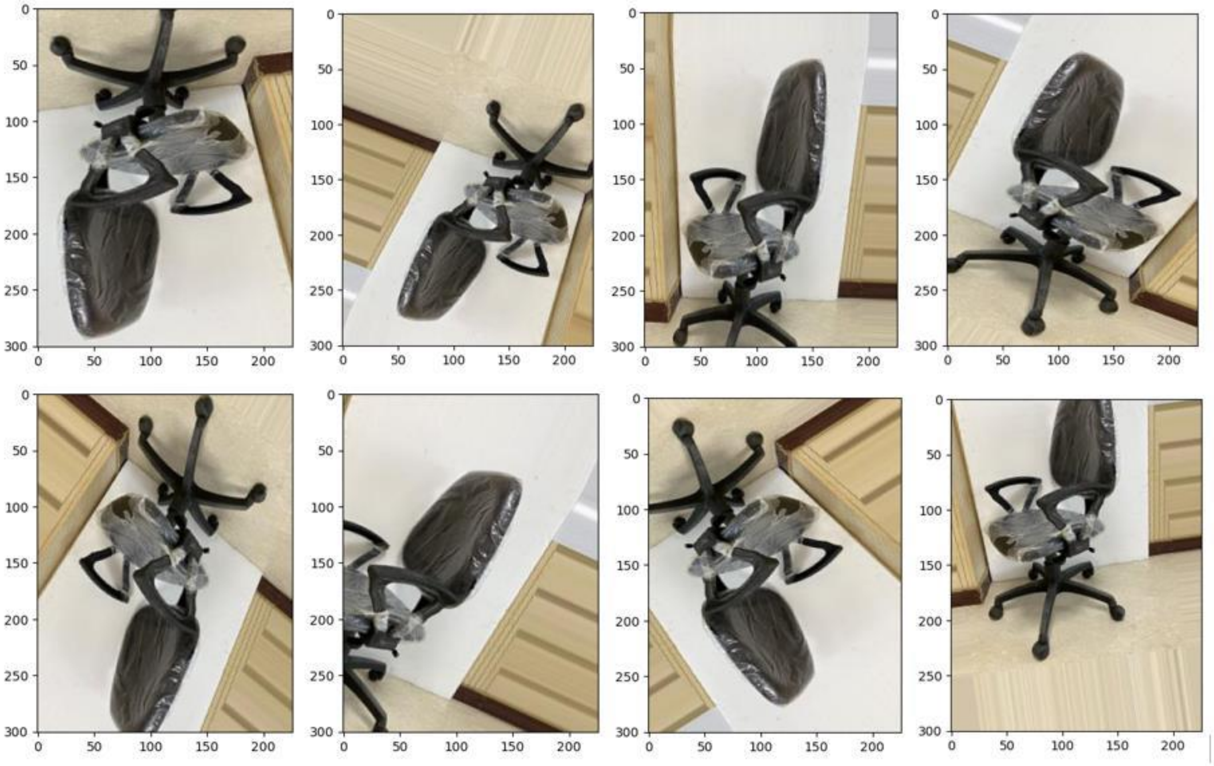
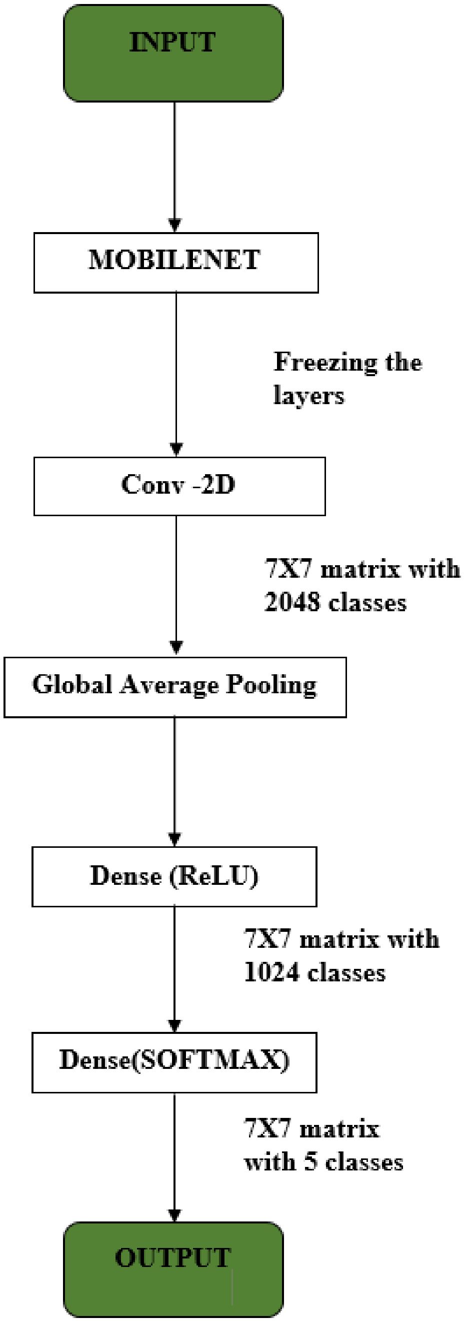


Fig 1.4 Data Augmentation

MobileNet Model

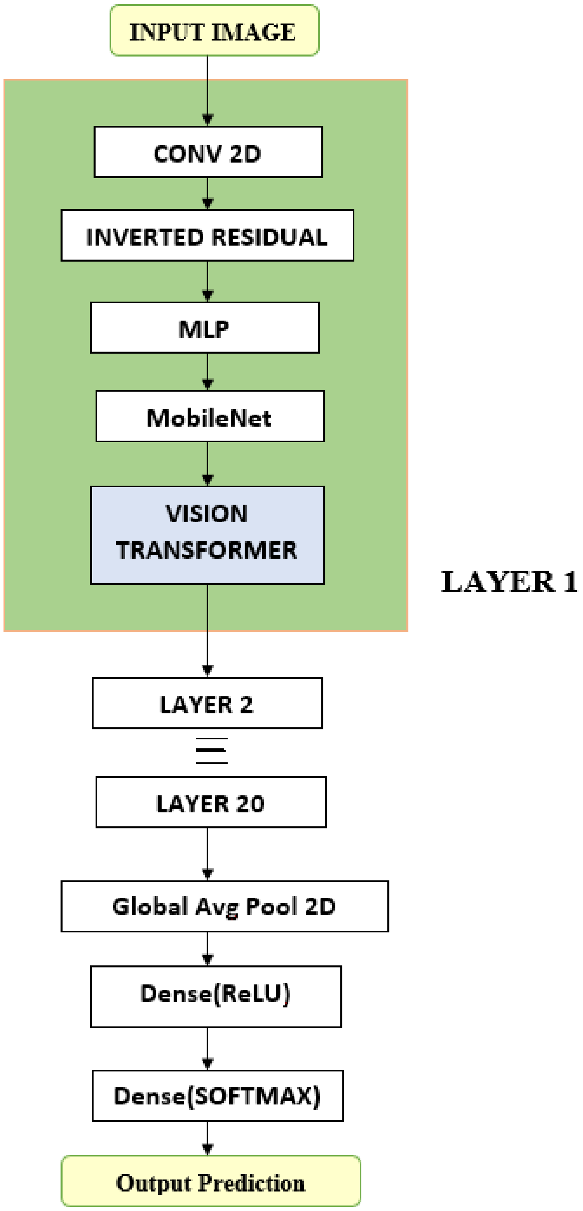
In the implementation of MobileNet, considering a pre-trained MobileNet model from tensorflow, and here simply training and testing of the algorithm takes place on a custom dataset. As this model is being used for a custom dataset, the top layer of the pretrained MobileNet model will be removed, with softmax being used as the classifier for activation. As this model is pre-trained on other dataset present on the tensorflow directories, so it is not required to use the pre-trained method. Because it is a very huge network, it is not required to train all the layers of the model. By simply freezing the trainable parameters of the original MobileNet model, the efficiency of the network increases. After freezing the trainable layers, it is required to add few layers on top of this model, which needs to be trained. In a deep neural network, ReLU[5] is typically employed as an activation function for the hidden layers. In order to do this, we leverage the activation of the penultimate layer in a neural network to learn the weight parameters of the ReLU classification layer by backpropagation [4].

Flow Chart of MobileNet



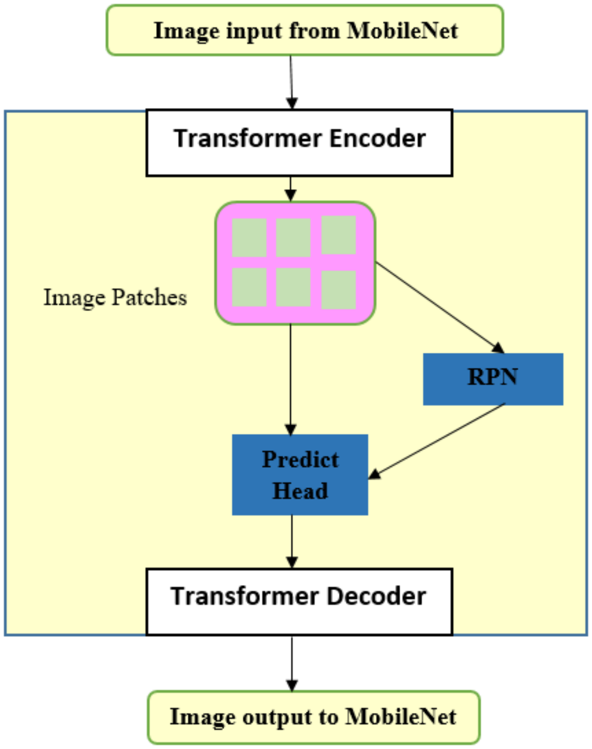
MobileNet ViT Model

Flow Chart of Proposed Method-



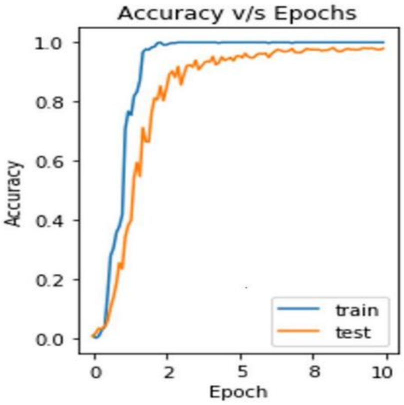
For the implementation of MobileNet with Vision Transformer, the major thing to be taken into consideration is Hyper parameters and image pixels when entering the ViT Block. Initially, for the input images of size 256x256, a rescaling of the images is done in order to reduce patch size for ViT block[6]. A 2D convolution layer is added to perform convolution twice on the same image. MobileNet uses depthwise convolution layer followed by batch normalization and fully convolution layers. After repeating the depthwise convolution nine times along with batch normalization and fully connected convolution layer, the layer normalization is done which makes the image suitable to enter the transformer block. The multihead attention layer is added which includes the vision transformer. All these layers are a part of single hidden layer, to make the network dense and more accurate these hidden layers are repeatedly connected twenty times, giving the network better accuracy. Once all the hidden layers are executed along with all the attention layers, to make the output more accurate, another layer of global average pooling is added, and a Dense layer after it is added, making the model suitable to be trained for five image classes. The Transformer Encode block incorporate a single image into image patches. After that, the picture patches are flattened. The flattened picture patches are used to construct "Patch Embeddings," which are lower-dimensional linear embeddings. The picture patch sequences are then given positional embeddings in order to preserve their positional information, which is essential for detecting anomalies on the frames. The trainable class encoding output from the MLP is simply piled on top of the trainable class output images. The classification process is completed at last [7].

II. Block Diagram of ViT Block



III. Results and Discussion

After the models has been trained on the custom dataset, to check how well the network has performed, there are certain tools that are used to measure the performance of the network, and these tools are called performance evaluation metrics. In this paper, using few common metrics for classification problems to gain valuable information about the performance of algorithms and to perform comparative analysis. These metrics include precision, recall [8], f1-score [9], accuracy [10] and confusion matrix [10].



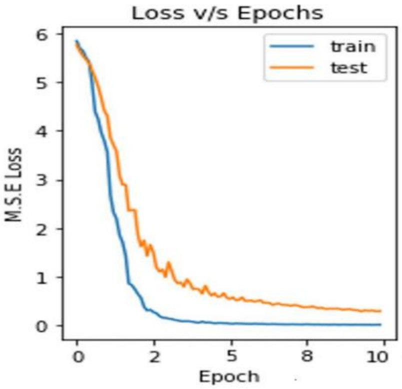


Fig 3.1 Accuracy vs Epochs graph for MobileNet

Fig 3.2 Loss vs Epochs graph for MobileNet

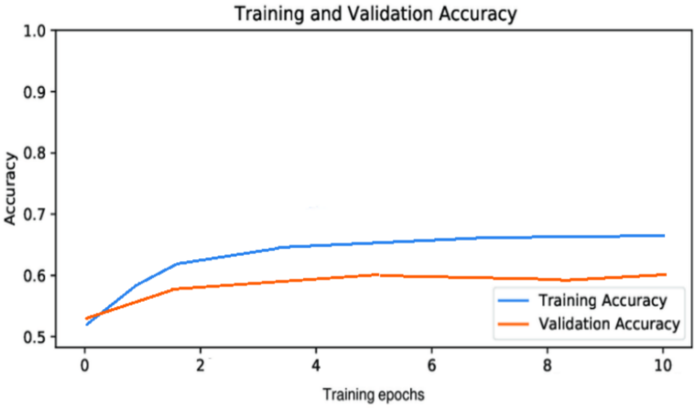


Fig 3.3 Training and Validation Accuracy for MobileNet

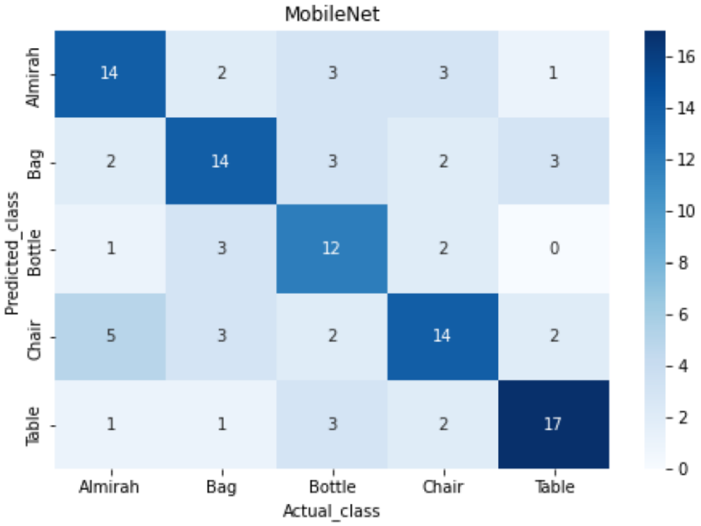
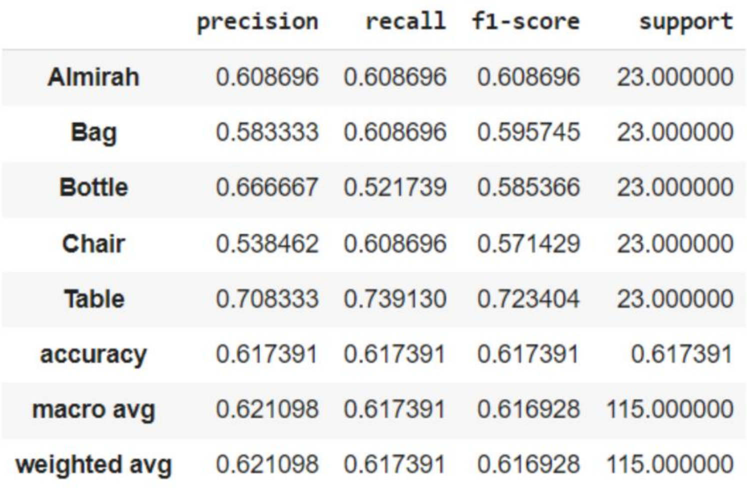


Fig 3.4 Confusion Matrix

Table 3.1 Evaluation Matrix for MobileNet model



	precision	recall	f1-score	support
Almirah	0.608696	0.608696	0.608696	23.000000
Bag	0.583333	0.608696	0.595745	23.000000
Bottle	0.666667	0.521739	0.585366	23.000000
Chair	0.538462	0.608696	0.571429	23.000000
Table	0.708333	0.739130	0.723404	23.000000
accuracy	0.617391	0.617391	0.617391	0.617391
macro avg	0.621098	0.617391	0.616928	115.000000
weighted avg	0.621098	0.617391	0.616928	115.000000

For MobileNet model the training and validation accuracy for all the epochs is shown in Fig 3.3. According the graph, the overall accuracy of the model is around 65% for Training and 55% for Validation, which makes the model less accurate. As shown in Fig 3.4, the confidence percentage by which the model shows that chair class objects in input is same as chair class objects is 94% confidence. Similarly, for Bag class the confidence value is 93%, but for other classes like, the confidence value for which the model tells that the almirah class in input is predicted as table class is 25%, making the model not a great model for detection. The parameters for evaluation are mentioned in table 3.1 with precision, accuracy, recall and f1-score. For all the classes in the dataset, the accuracy is observed to be 61.7% using conventional MobileNet model on the custom dataset.

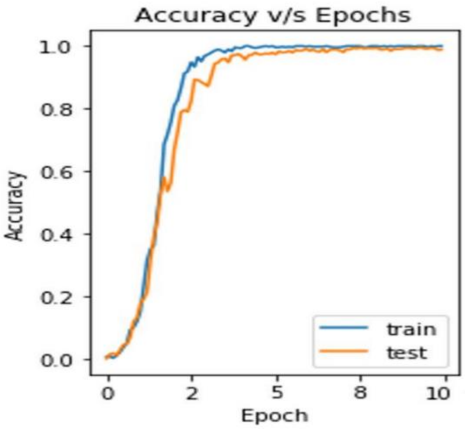


Fig 3.6 Accuracy vs Epochs graph for MobileNet ViT

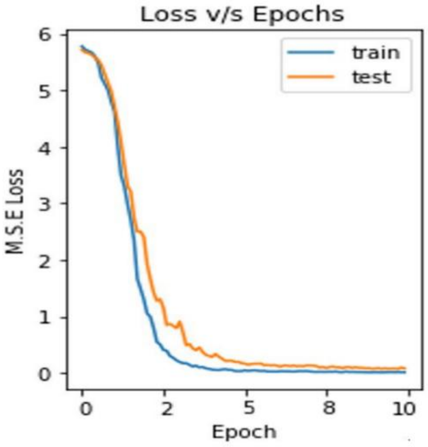


Fig 3.7 Loss vs Epochs graph

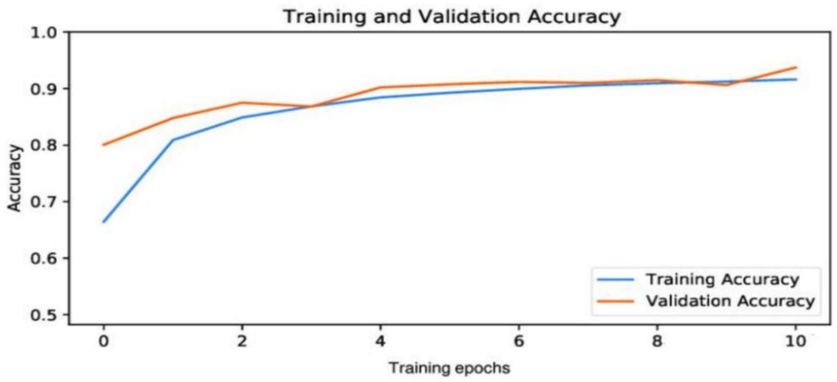
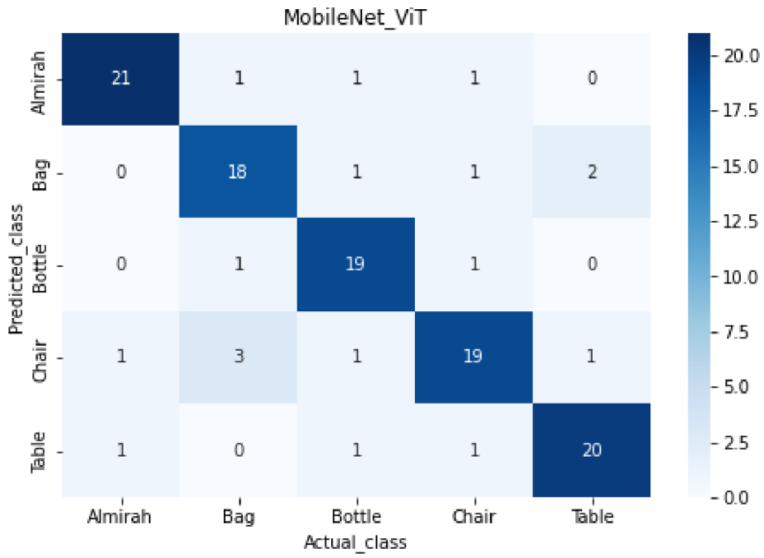


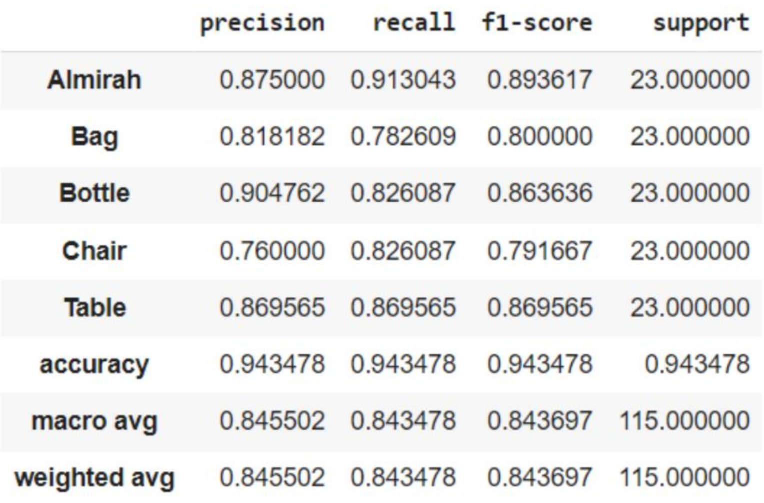
Fig 3.8 Training and Validation Accuracy for MobileNet ViT



	Almirah	Bag	Bottle	Chair	Table
Almirah	21	1	1	1	0
Bag	0	18	1	1	2
Bottle	0	1	19	1	0
Chair	1	3	1	19	1
Table	1	0	1	1	20

Fig 3.9 Confusion Matrix for MobileNet ViT model

Table 3.2 Evaluation Matrix for MobileNet ViT



	precision	recall	f1-score	support
Almirah	0.875000	0.913043	0.893617	23.000000
Bag	0.818182	0.782609	0.800000	23.000000
Bottle	0.904762	0.826087	0.863636	23.000000
Chair	0.760000	0.826087	0.791667	23.000000
Table	0.869565	0.869565	0.869565	23.000000
accuracy	0.943478	0.943478	0.943478	0.943478
macro avg	0.845502	0.843478	0.843697	115.000000
weighted avg	0.845502	0.843478	0.843697	115.000000

According to the graph shown in figure 3.8, the training and validation accuracies are overlapping over the period of 10 epochs. Also the model has given 95% accuracy for both the training and validation datasets. The different in accuracies is the least in this model. This increase in accuracy, results in a network with less error, which is done by adding layers of Vision Transformer in the MobileNet network. The Confusion matrix shows for an input image of class chair, the confidence by which the predicted image belong to chair class is 94%. Similarly, for an input image of class table, the confidence of the network that predicted object also belongs to the same class is 96%. The values are higher for the same actual and predicted classes. For other predicted classes which are not same as actual classes, the confidence values are least, minimizing the error index of this network. The parameters for evaluation are mentioned below in table 3.2 with precision, accuracy, recall and f1-score. For all the classes in the dataset, the accuracy is observed to be 94.3% using Vision Transformer with MobileNet

IV. Conclusion

After observing the accuracy value in table 3.1, it can be concluded that the overall accuracy of conventional MobileNet model on custom dataset is 61.7%, whereas from table 3.4 the accuracy of MobileNet model with Vision Transformer reaches 94.3% for the same custom dataset. Thus making MobileNet ViT a suitable model for object detection for mobile robot applications. A conventional MobileNet used for object detection becomes a deep network with multiple layers, making it more complex and time consuming. Adding a layer of Vision Transformer in the MobileNet algorithm, gives a vast difference in the network output. The vision transformer which is based on the concept of self-attention, outperform existing transformer models while being faster than most competitive CNNs like MobileNet, shown by extensive experiments on object detection in Laboratory.

The use of Vision Transformer to counter convolutional network helped the network use "self-attention" to focus on key object segments in the image, reduce computationally intensive operations, and provide better results as the positional deployment helped the network learn about the various image segments and the relevance of their position, which was not possible in convolutional network. The transformer network also has a better operational capability than the conventional model.

References

1. J. Beal et al. *Toward transformer-based object detection*. arXiv:2012.09958, 2020

2. Pranav Adarsh, Pratibha Rathi, and Manoj Kumar. Yolo v3-tiny: Object detection and recognition using one stage improved model. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), pages 687–694. IEEE, 2020.

3. https://jonathan-hui.medium.com/ssd-object-detection-single-shot-multibox-detector-for-real-time-processing-9bd8deac0e06

4. Deep Learning using Rectified Linear Units (ReLU) Abien Fred M. Agarap, 2019.

5. S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015.\

6. Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In Advances in Neural Information Processing Systems. 577–585.

7. 42. MobileViT : Light-weight, General-purpose, and Mobile-friendly Vision Transformer Sachin Mehta, Mohammad Rastegari 2021 ICLR 2022 Conference

8. D. M. Powers, "Evaluation: from precision, recall and Fmeasure to ROC, informedness, markedness and correlation," Journal of machine learning research, vol. 2, 2011, pp. 37– 63.

9. Y. Sasaki, "The truth of the F-measure," Teach Tutor mater, vol. 1, no. 5, 2007, pp. 1-5.

10. Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification Meysam Vakili, Mohammad Ghamsari and Masoumeh Rezaei 2020.