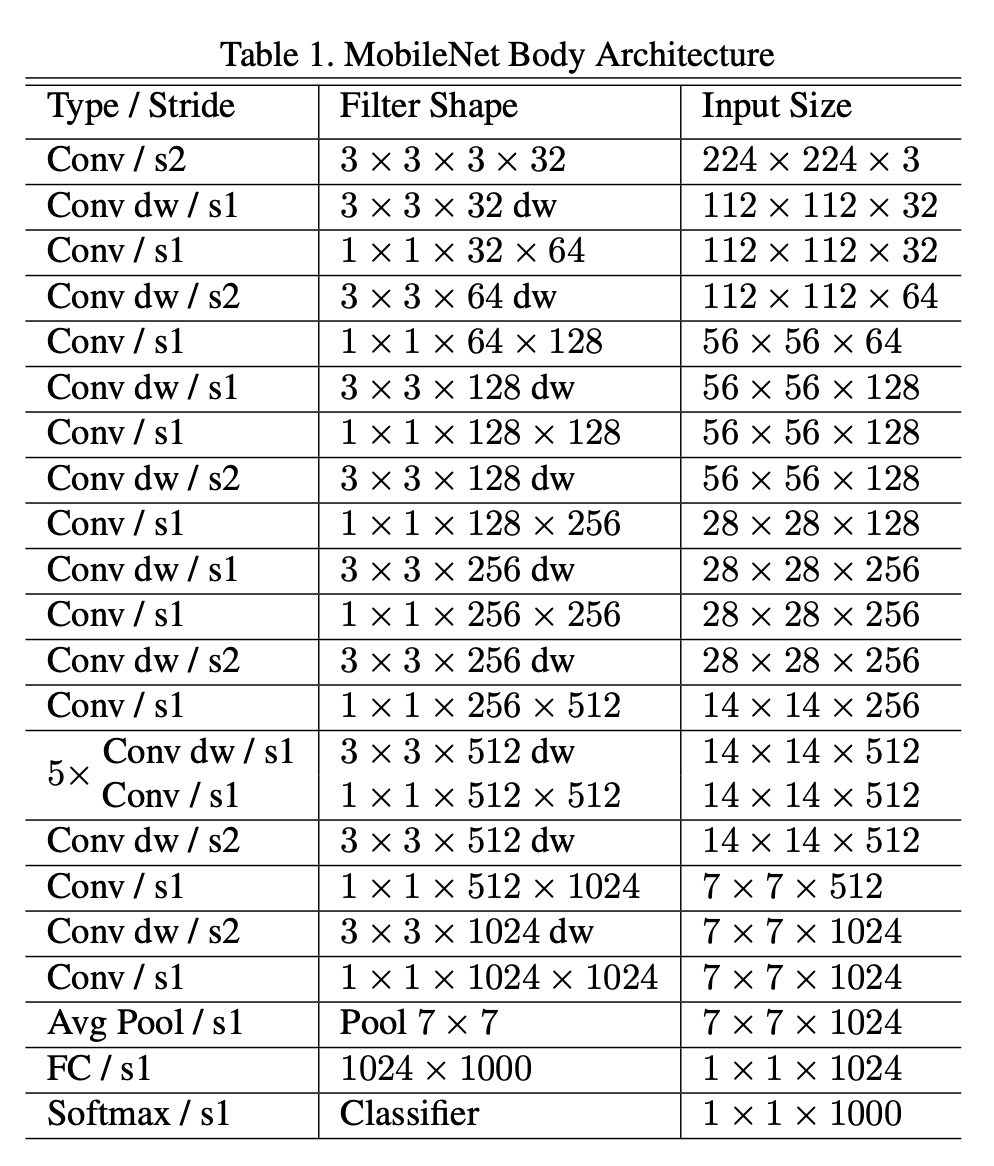
**2.Project Hypothesis and definition**

* Understanding the mobilenet v1 architecture
* Explaining point wise ,depth wise,normal convolution and architecture
* Implementation of the architecture and getting the parameters and accuracy (base model)
* Reading different papers to reduce the number of parameters of the mobilenetv1

The main task is to deploy the machine learning algorithms in FPGA .As we know hardware chips have very low memory to store the weights of the model.My research was to reduce the number of parameters of the mobilenetv1 network .

## Understanding the mobile netv1 architecture:

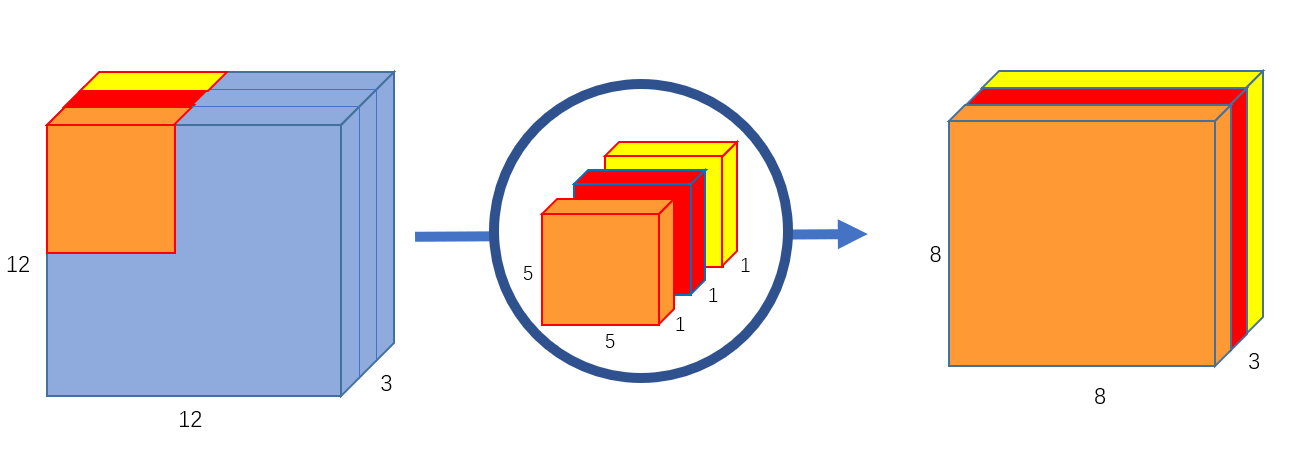
MobileNet is one of the successful model and widely used in mobile and embedded systems



The mobilenetv1 contains a set of convolutions with different filter size and stride along with depthwise convolution and point wise convolution in order to reduce the parameters of convolution and can use instead of convolution

## Explaining point wise ,depth wise,normal convolution and architecture

### 1.Depthwise separable

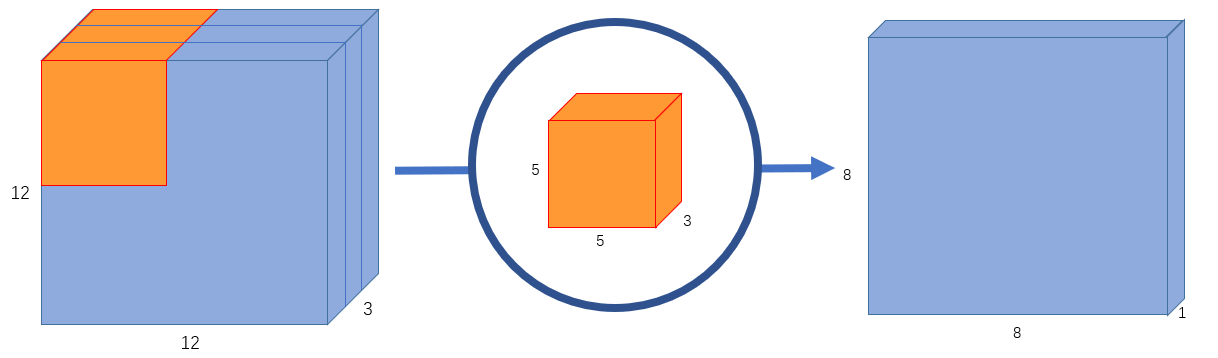


Depth Wise separable is nothing but the convolution Which is done in channel wise instead of the whole block .The result obtained is given to the point wise convolution in order to get the result similar with the normal convolution

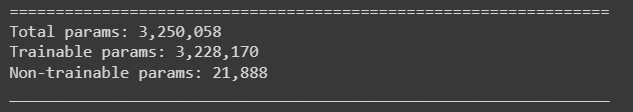
### 2.pointwise convolution

Pointwise convolution is applied on depthwise convolution in order to aggregate all the channels that we got from depthwise convolution. The pointwise operation is done with filter with channels of the depthwise

### 3.Normal Convolution



A Normal convolution is done on an image with a filter ,Each filter is convolved on the image to get the output



## Implementation of the architecture and getting the parameters and accuracy (base model)

The implementation of mobilenetV1 can be found at the [colab link](https://colab.research.google.com/drive/1aek1xR4lT55baW2wTJhP28Oz_sPrwBdQ?usp=sharing)

## Reading different papers to reduce the number of parameters of the mobilenetv1

Different sources are referred to reduce the number of parameters of the convolution layer.Each paper reduces the parameters differently here i am writing the definitions of the operators used in the code

#### Group convolution:

Here the filters are divided into groups and the same group kernels are convolved so there is a reduction in the number of parameters as we are using less number of filters .If there are 3 groups given as the hyper parameter of the convolution then we are reducing the number of parameters

#### Predefined kernel:

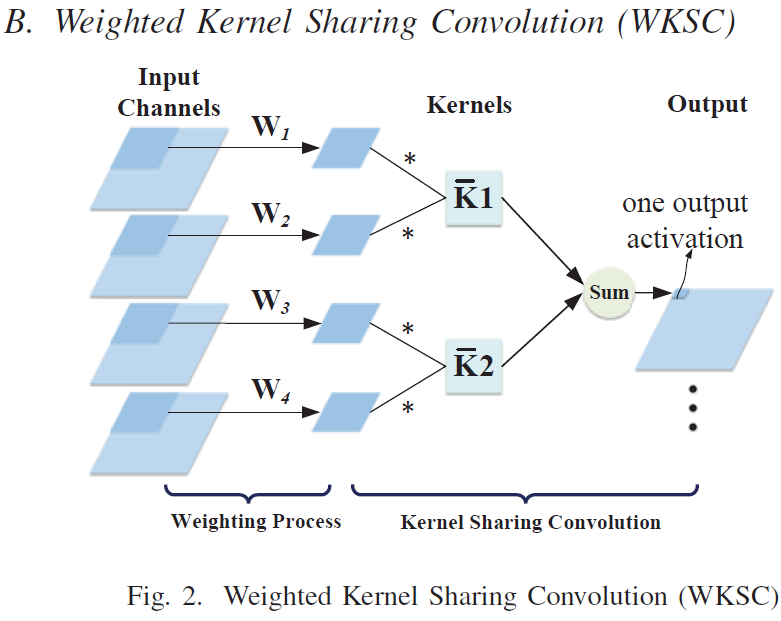
Here we are using the kernel weights which are predefined which are not not used in training the.These are kept as constants both in forward and backward pass

#### High level autocorrelation:

In this we are using High level autocorrelation ,where we predefine 25 kernels and we determine the 25 kernels using High level autocorrelation .In my code we are using 5 orders of HLAC which is like different 0,1 arranged in matrix for the five times so that we get the correlated values from the given value which we are using it as the constant kernels

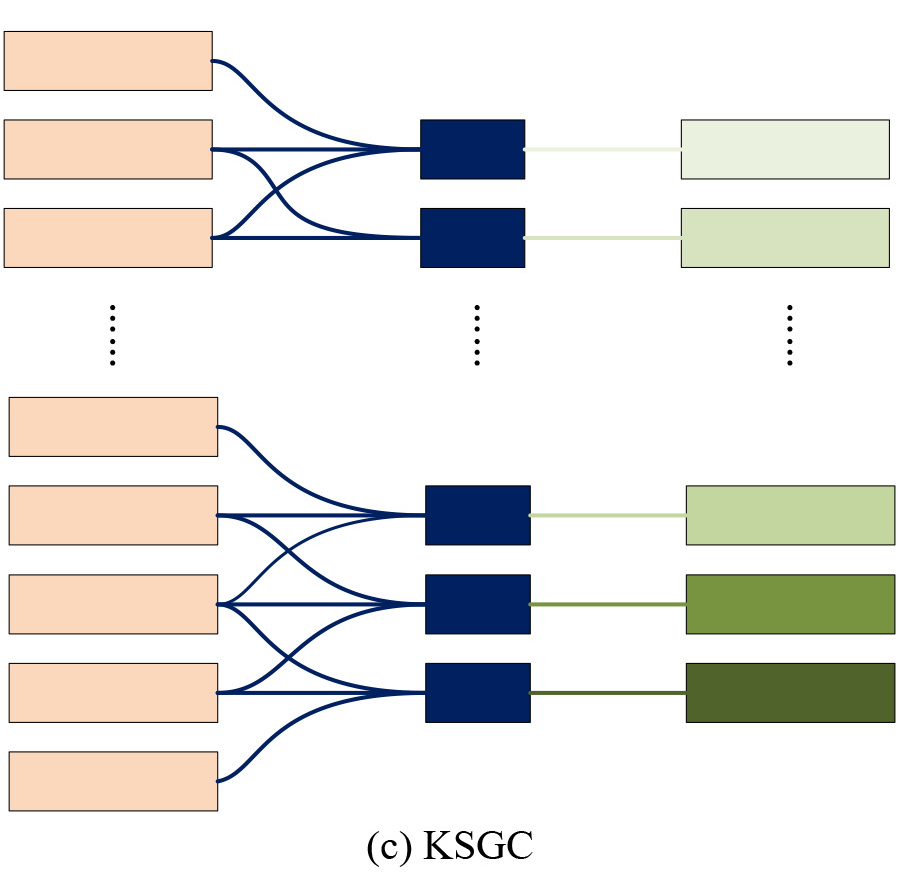
#### Weighted kernel sharing convolution:

In this convolution we apply the weights to the input channel this process is called weighing process .after applying the weights the kernel are shared



#### Kernel sharing group convolution

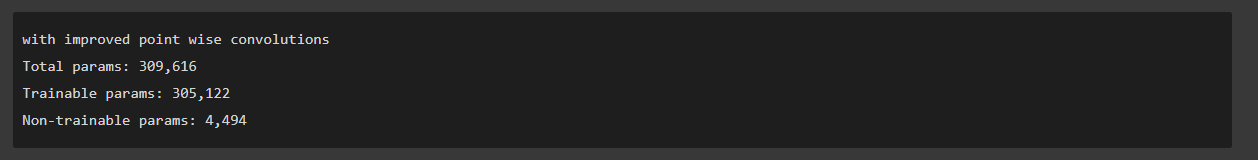
In a kernel sharing group ,different groups share the same kernel so if there are n filters given then based on the group sharing the number of filters used will be reduced .



#### Improved point wise convolution:

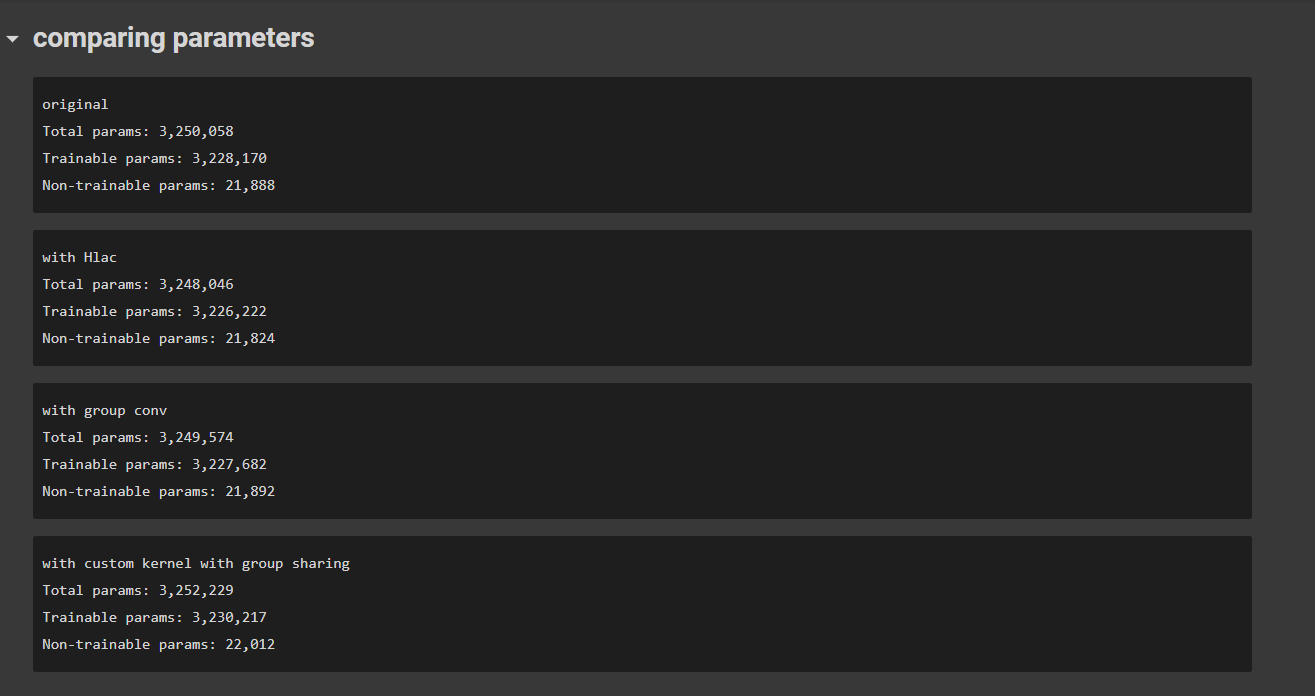
* In this we try to replace the point wise convolution as these layers contribute more parameters to the mobilenet.The total no of computations that we see in pointwise convolution is .we need to decompose this inorder do that we will make a 2d seed which is randomly generated from -1 to 1 which is of shape of the channels as depthwise convolution in the previous layer
* After that we apply 1D convolutions with k kernels of size |k| we will be getting as the output
* After that we apply the concatenation between the and the pointwise operation with the output of the depthwise convolution which can be used to reduce the parameters.
* In this paper they used H=14,W=14,|k|=9,the parameters can be reduced by 4.6%

Results for Improved Pointwise convolution



Had a problem with reshaping and applying the conv2d and activation need to check the code and update the reshape .

#### All parameters of the different implementations:



## Draw Back

* All Group convolution,Predefined kernel,High Level Auto correlation,Weighted kernel sharing convolution,kernel sharing group convolution are used to **reduce** the parameters of the **conv2D** but we are having only one conv2d .so it doesn't effect the total number of parameters
* In order to reduce the parameters of the pointwise convolution we are using the **Improved point wise convolution.**

### 