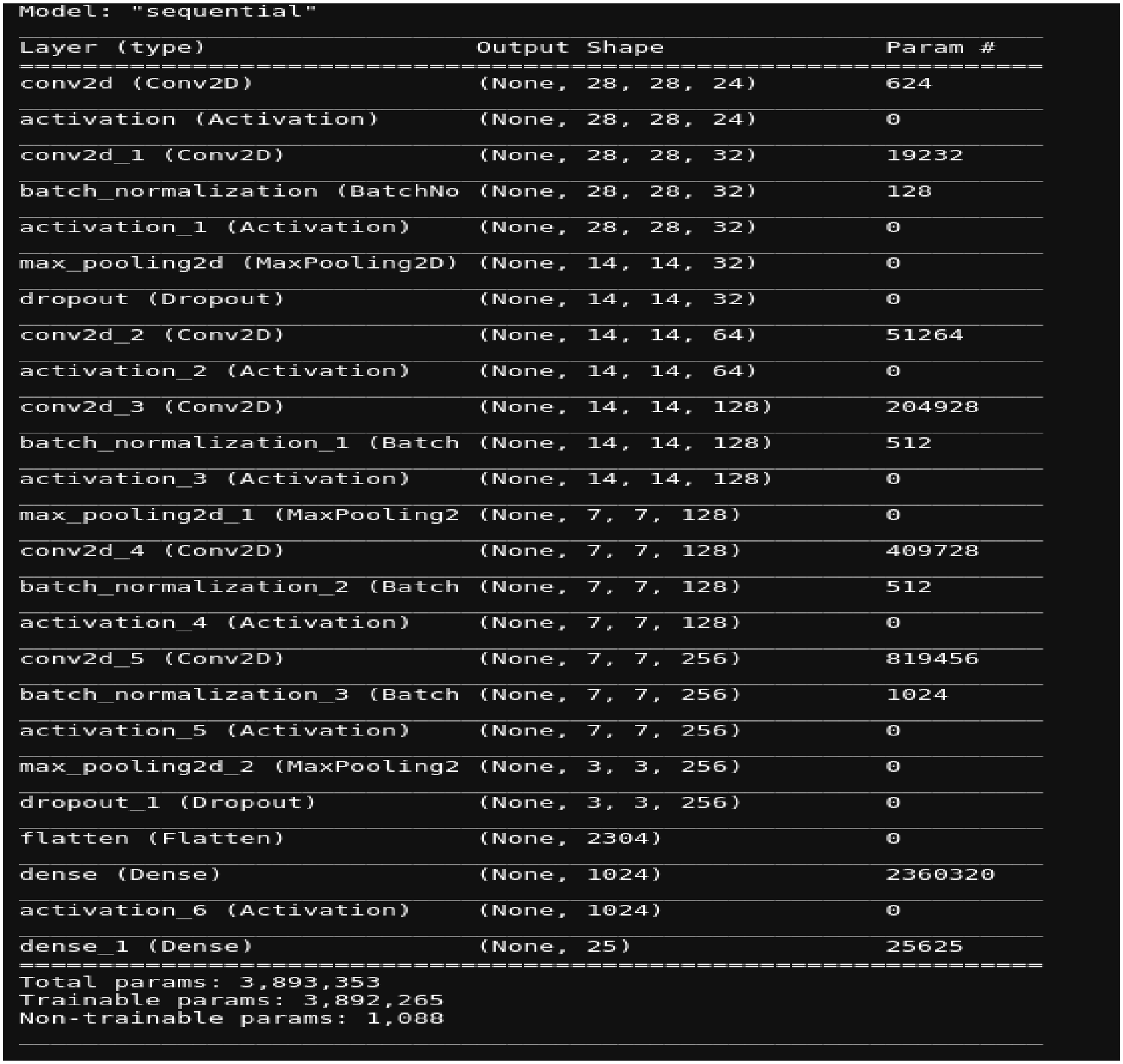
Question 1: All observations and calculation are done in notebook (question1.ipynb)

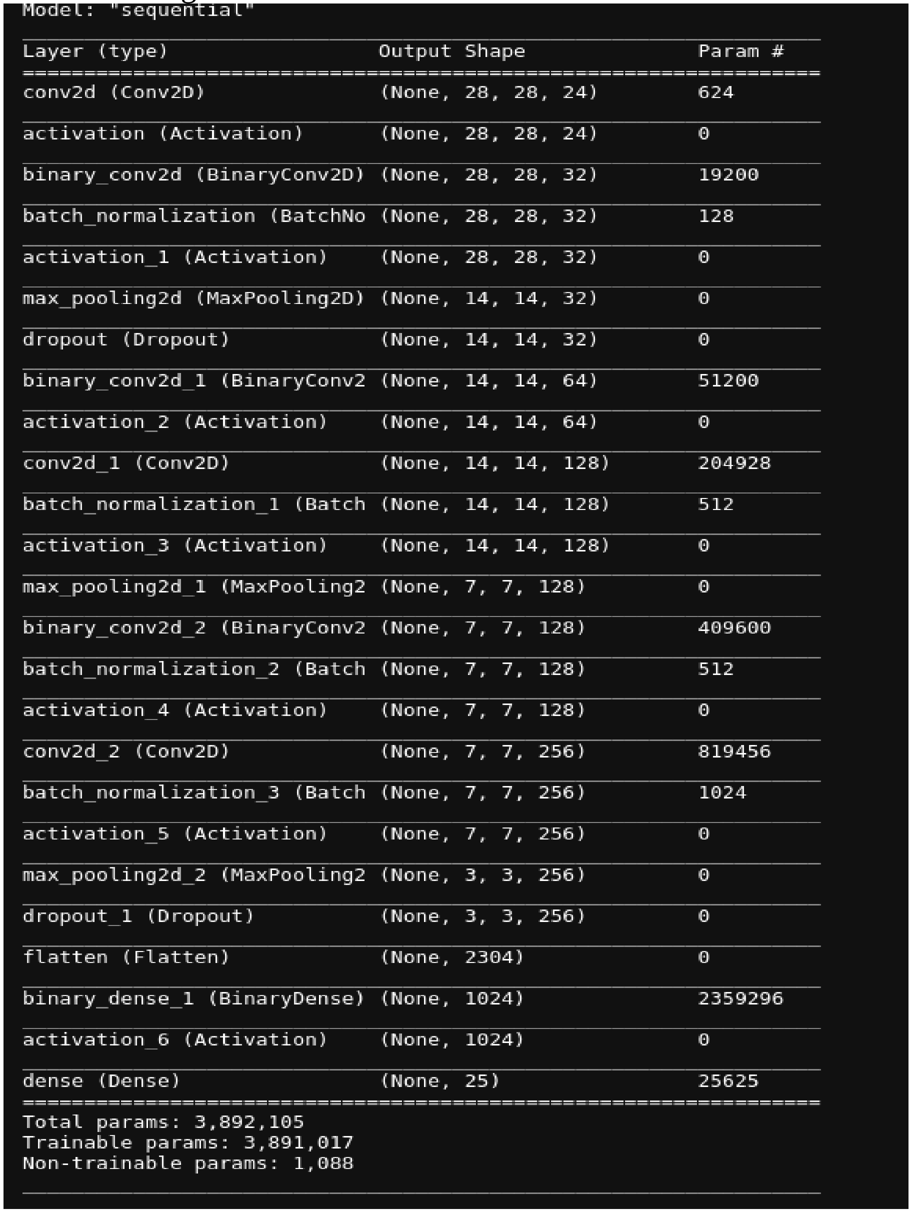
Question: 2
Base model calculations :



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 28, 28, 24)        624
_________________________________________________________________
activation (Activation)      (None, 28, 28, 24)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 28, 28, 32)        19232
_________________________________________________________________
batch_normalization (BatchNo (None, 28, 28, 32)        128
_________________________________________________________________
activation_1 (Activation)    (None, 28, 28, 32)        0
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 14, 14, 32)        0
_________________________________________________________________
dropout (Dropout)            (None, 14, 14, 32)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 14, 14, 64)        51264
_________________________________________________________________
activation_2 (Activation)    (None, 14, 14, 64)        0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 14, 14, 128)       204928
_________________________________________________________________
batch_normalization_1 (Batch (None, 14, 14, 128)       512
_________________________________________________________________
activation_3 (Activation)    (None, 14, 14, 128)       0
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 7, 7, 128)         0
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 7, 7, 128)         409728
_________________________________________________________________
batch_normalization_2 (Batch (None, 7, 7, 128)         512
_________________________________________________________________
activation_4 (Activation)    (None, 7, 7, 128)         0
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 7, 7, 256)         819456
_________________________________________________________________
batch_normalization_3 (Batch (None, 7, 7, 256)         1024
_________________________________________________________________
activation_5 (Activation)    (None, 7, 7, 256)         0
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 3, 3, 256)         0
_________________________________________________________________
dropout_1 (Dropout)          (None, 3, 3, 256)         0
_________________________________________________________________
flatten (Flatten)            (None, 2304)              0
_________________________________________________________________
dense (Dense)                (None, 1024)              2360320
_________________________________________________________________
activation_6 (Activation)    (None, 1024)              0
_________________________________________________________________
dense_1 (Dense)              (None, 25)                25625
=================================================================
Total params: 3,893,353
Trainable params: 3,892,265
Non-trainable params: 1,088
_________________________________________________________________
```

Accuracy = 97.21%
Total params = 3,893,353
Memory foot print = total params * 32 = 124587296 bits = 15.573412 Mb

2a. All calculations are done by Assuming binary layer uses 2 bits, convo2d layers and all other layers uses 32 bits for storage.

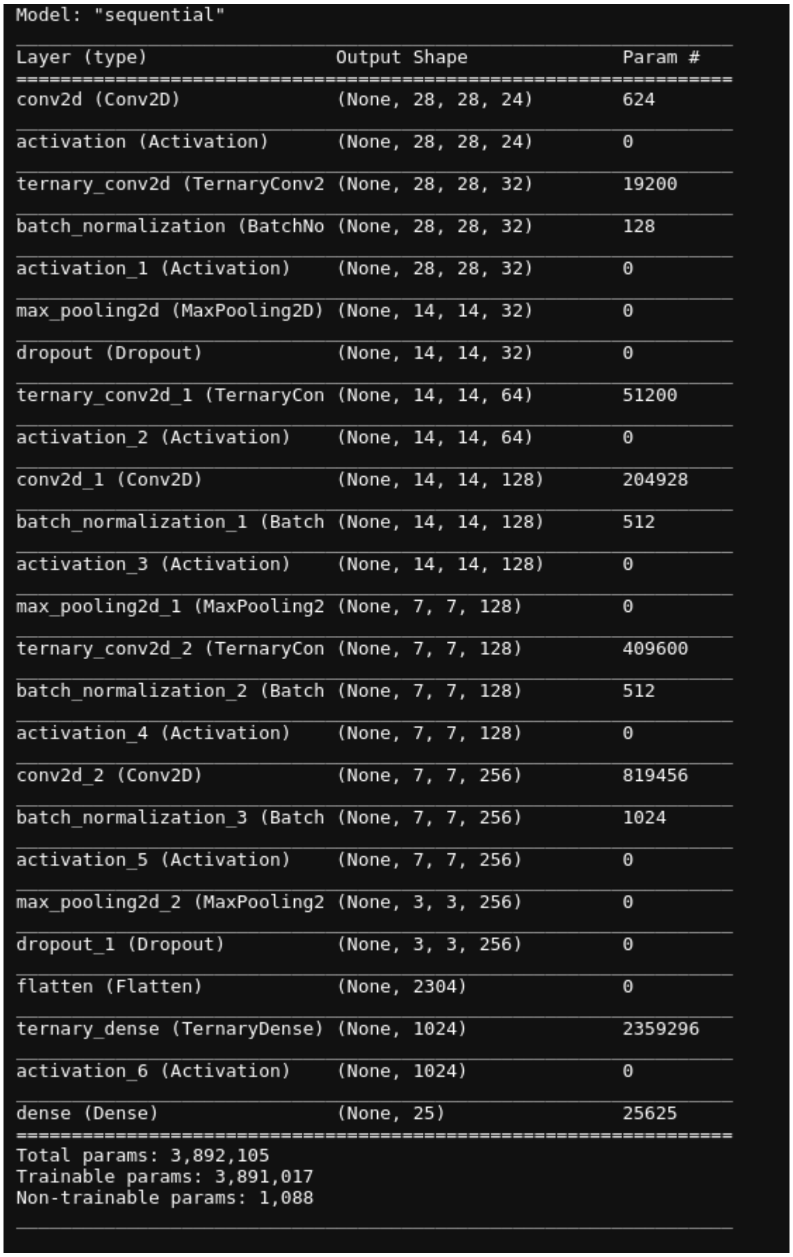


```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 28, 28, 24)        624
_________________________________________________________________
activation (Activation)      (None, 28, 28, 24)        0
_________________________________________________________________
binary_conv2d (BinaryConv2D) (None, 28, 28, 32)        19200
_________________________________________________________________
batch_normalization (BatchNo (None, 28, 28, 32)        128
_________________________________________________________________
activation_1 (Activation)    (None, 28, 28, 32)        0
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 14, 14, 32)        0
_________________________________________________________________
dropout (Dropout)            (None, 14, 14, 32)        0
_________________________________________________________________
binary_conv2d_1 (BinaryConv2 (None, 14, 14, 64)        51200
_________________________________________________________________
activation_2 (Activation)    (None, 14, 14, 64)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 14, 14, 128)       204928
_________________________________________________________________
batch_normalization_1 (Batch (None, 14, 14, 128)       512
_________________________________________________________________
activation_3 (Activation)    (None, 14, 14, 128)       0
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 7, 7, 128)         0
_________________________________________________________________
binary_conv2d_2 (BinaryConv2 (None, 7, 7, 128)         409600
_________________________________________________________________
batch_normalization_2 (Batch (None, 7, 7, 128)         512
_________________________________________________________________
activation_4 (Activation)    (None, 7, 7, 128)         0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 7, 7, 256)         819456
_________________________________________________________________
batch_normalization_3 (Batch (None, 7, 7, 256)         1024
_________________________________________________________________
activation_5 (Activation)    (None, 7, 7, 256)         0
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 3, 3, 256)         0
_________________________________________________________________
dropout_1 (Dropout)          (None, 3, 3, 256)         0
_________________________________________________________________
flatten (Flatten)            (None, 2304)              0
_________________________________________________________________
binary_dense_1 (BinaryDense) (None, 1024)              2359296
_________________________________________________________________
activation_6 (Activation)    (None, 1024)              0
_________________________________________________________________
dense (Dense)                (None, 25)                25625
=================================================================
Total params: 3,892,105
Trainable params: 3,891,017
Non-trainable params: 1,088
_________________________________________________________________
```

Accuracy = 91.46%
Total Binary layer params = 2839296
Binary layer memory foot print = 2839296 * 2 = 5678592
Other layers memory foot print = 3,892,105 - 2839296 =1052809 *32 = 33689888
Total memory foot print with binary layers = 39368480 bits = 4.92106 Mb
Reduction in memory foot print = ((15.573412 - 4.92106) / 15.573412) *100 = 68.40 %
Total memory Reduction = 68.40 %

There is a 68.40 % reduction in memory food print with binary layer compared to the original model . The model accuracy is dropped by 5.75% which is comparatively a good trade of between accuracy and size of the model.

2b. All calculations are done by Assuming Ternary layer uses 3 bits, convo2d layers and all other layers uses 32 bits for storage.



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 28, 28, 24)        624
_________________________________________________________________
activation (Activation)      (None, 28, 28, 24)        0
_________________________________________________________________
ternary_conv2d (TernaryConv2 (None, 28, 28, 32)        19200
_________________________________________________________________
batch_normalization (BatchNo (None, 28, 28, 32)        128
_________________________________________________________________
activation_1 (Activation)    (None, 28, 28, 32)        0
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 14, 14, 32)        0
_________________________________________________________________
dropout (Dropout)            (None, 14, 14, 32)        0
_________________________________________________________________
ternary_conv2d_1 (TernaryCon (None, 14, 14, 64)        51200
_________________________________________________________________
activation_2 (Activation)    (None, 14, 14, 64)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 14, 14, 128)       204928
_________________________________________________________________
batch_normalization_1 (Batch (None, 14, 14, 128)       512
_________________________________________________________________
activation_3 (Activation)    (None, 14, 14, 128)       0
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 7, 7, 128)         0
_________________________________________________________________
ternary_conv2d_2 (TernaryCon (None, 7, 7, 128)         409600
_________________________________________________________________
batch_normalization_2 (Batch (None, 7, 7, 128)         512
_________________________________________________________________
activation_4 (Activation)    (None, 7, 7, 128)         0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 7, 7, 256)         819456
_________________________________________________________________
batch_normalization_3 (Batch (None, 7, 7, 256)         1024
_________________________________________________________________
activation_5 (Activation)    (None, 7, 7, 256)         0
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 3, 3, 256)         0
_________________________________________________________________
dropout_1 (Dropout)          (None, 3, 3, 256)         0
_________________________________________________________________
flatten (Flatten)            (None, 2304)              0
_________________________________________________________________
ternary_dense (TernaryDense) (None, 1024)              2359296
_________________________________________________________________
activation_6 (Activation)    (None, 1024)              0
_________________________________________________________________
dense (Dense)                (None, 25)                25625
=================================================================
Total params: 3,892,105
Trainable params: 3,891,017
Non-trainable params: 1,088
_________________________________________________________________
```

Accuracy = 96.05%
Total Ternary layer params = 2839296
Ternary layer memory foot print = 2839296 * 3 = 8517888
Other layers memory foot print = 3,892,105 - 2839296 =1052809 *32 = 33689888
Total memory foot print with Ternary layers = 42207776 bits = 5.275972 Mb
Reduction in memory foot print = ((15.573412 - 5.275972) / 15.573412) *100 = 66.12 %
Total memory Reduction = 66.12 %

There is a 66.12 % reduction in memory food print with ternary layer compared to the original model . The model accuracy is dropped by 1.16 % which is comparatively a good trade of between accuracy and size of the model.

Q4.

Extra credit:

Model	Size	Model size
Base model	0.6866	5.77 Mb
Purned	0.6920	2.01 Mb
Purned and Quantized	0.6811	0.52 Mb