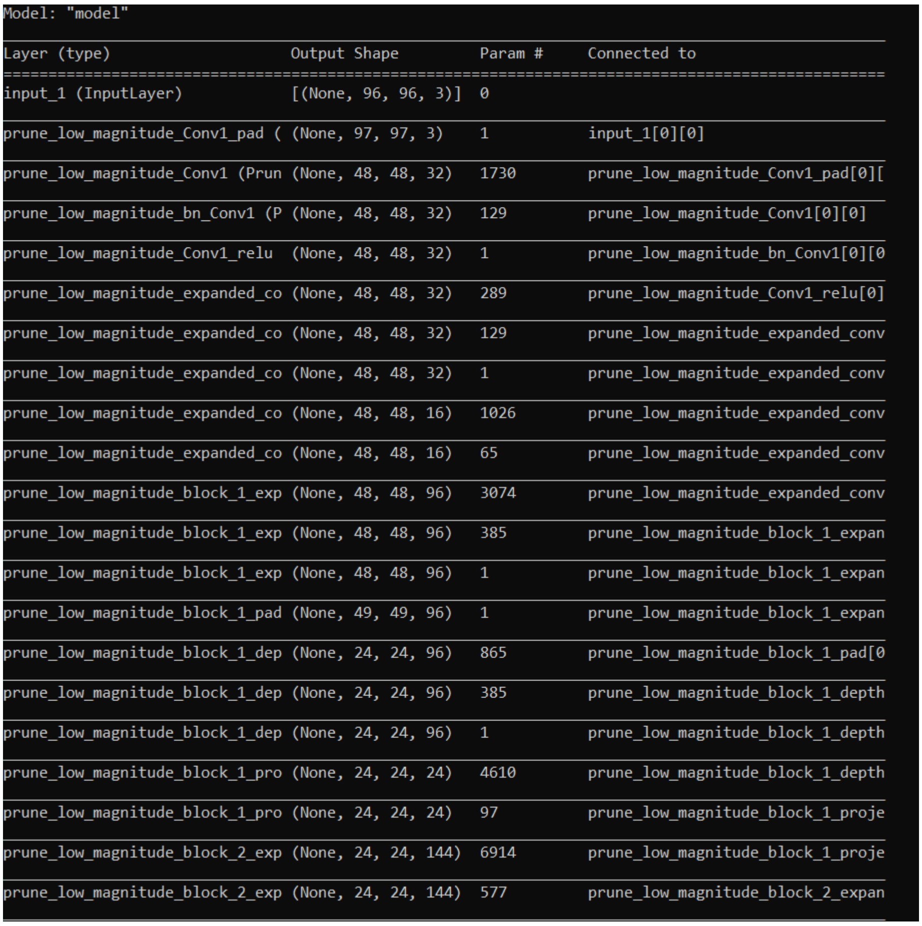
葉偉良 108065511

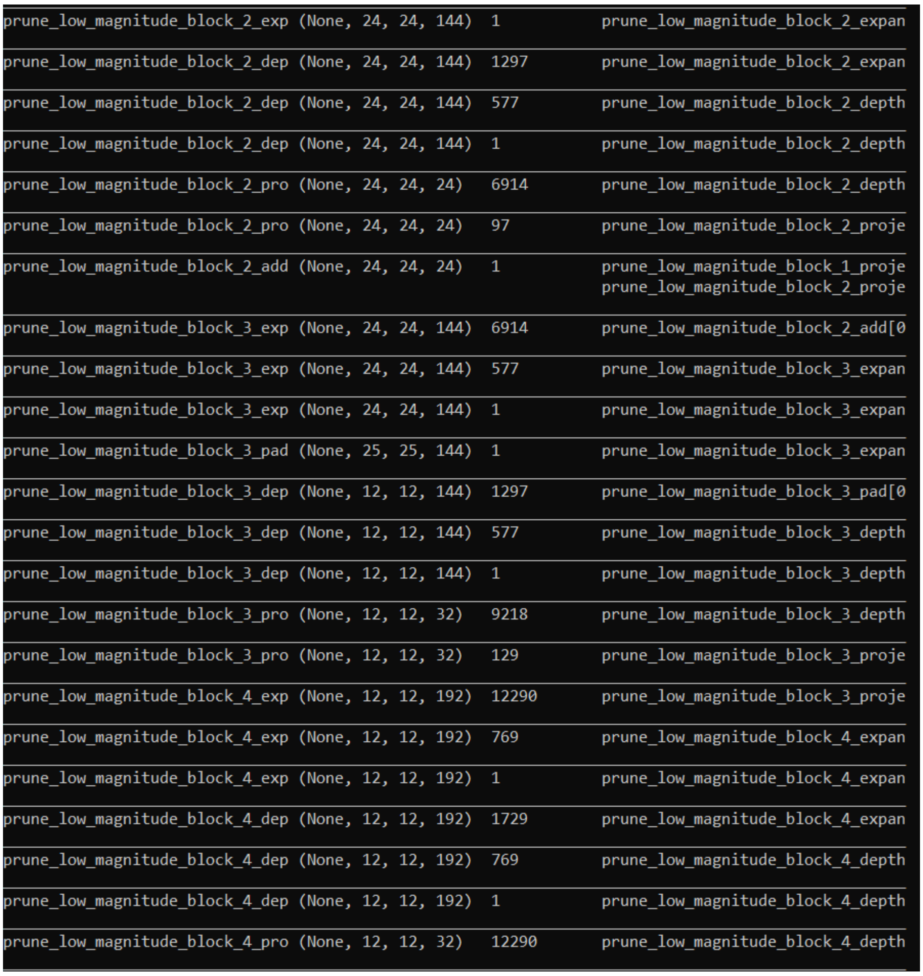
Report

1. Task 1 (pruning.py)

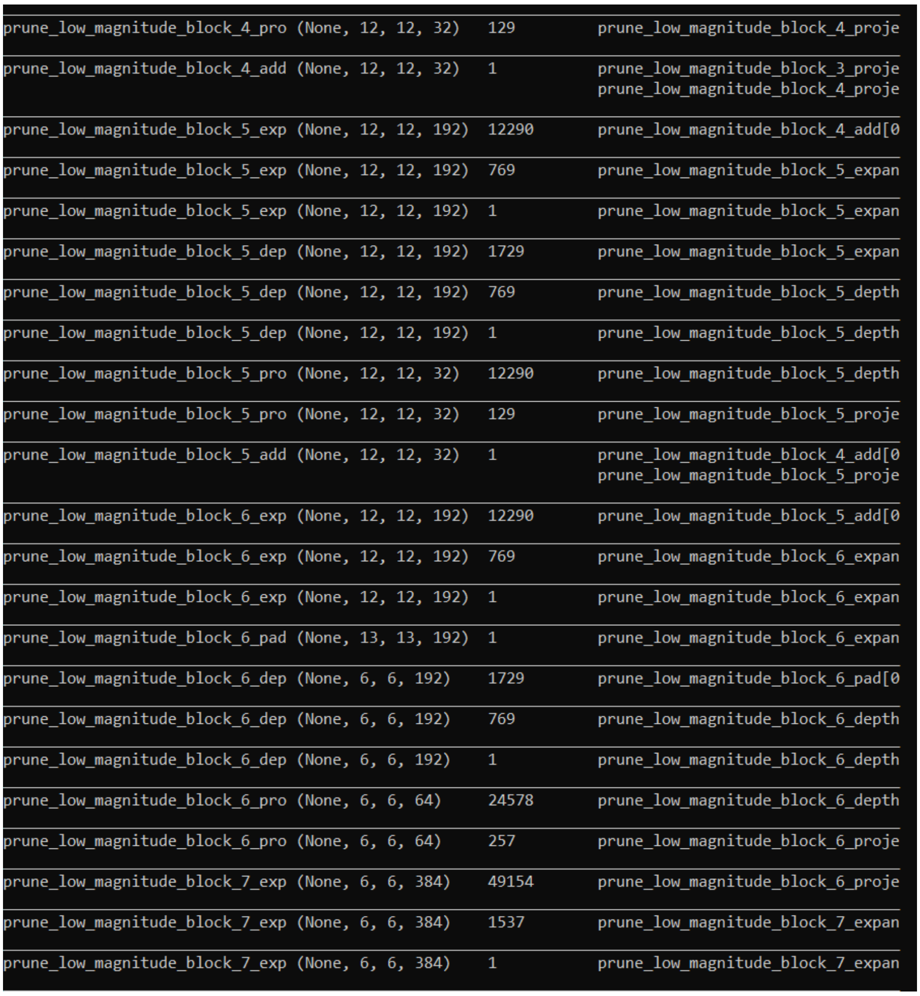
1.1 Model summary



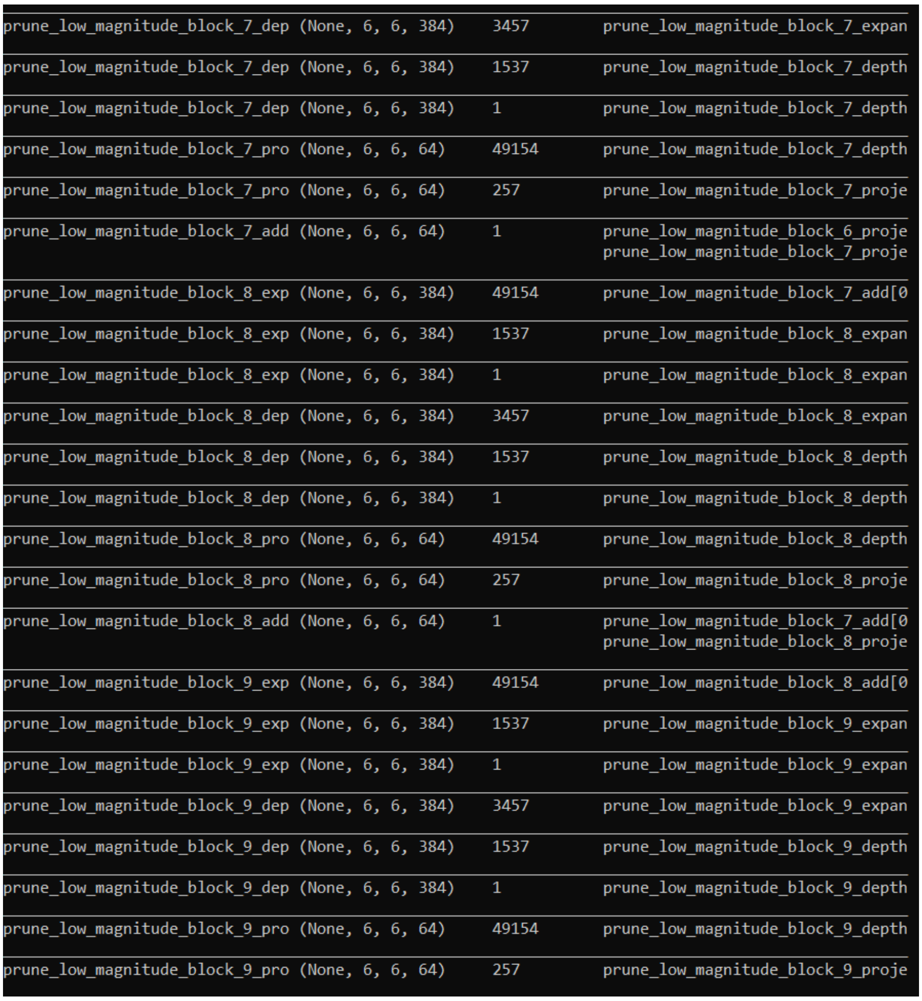
```
Model: "model"
__________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
input_1 (InputLayer)            [(None, 96, 96, 3)]  0
__________________________________________________________________________________________________
prune_low_magnitude_Conv1_pad ( (None, 97, 97, 3)    1           input_1[0][0]
__________________________________________________________________________________________________
prune_low_magnitude_Conv1 (Prun (None, 48, 48, 32)   1730        prune_low_magnitude_Conv1_pad[0][
__________________________________________________________________________________________________
prune_low_magnitude_bn_Conv1 (P (None, 48, 48, 32)   129         prune_low_magnitude_Conv1[0][0]
__________________________________________________________________________________________________
prune_low_magnitude_Conv1_relu  (None, 48, 48, 32)   1           prune_low_magnitude_bn_Conv1[0][0
__________________________________________________________________________________________________
prune_low_magnitude_expanded_co (None, 48, 48, 32)   289         prune_low_magnitude_Conv1_relu[0]
__________________________________________________________________________________________________
prune_low_magnitude_expanded_co (None, 48, 48, 32)   129         prune_low_magnitude_expanded_conv
__________________________________________________________________________________________________
prune_low_magnitude_expanded_co (None, 48, 48, 32)   1           prune_low_magnitude_expanded_conv
__________________________________________________________________________________________________
prune_low_magnitude_expanded_co (None, 48, 48, 16)   1026        prune_low_magnitude_expanded_conv
__________________________________________________________________________________________________
prune_low_magnitude_expanded_co (None, 48, 48, 16)   65          prune_low_magnitude_expanded_conv
__________________________________________________________________________________________________
prune_low_magnitude_block_1_exp (None, 48, 48, 96)   3074        prune_low_magnitude_expanded_conv
__________________________________________________________________________________________________
prune_low_magnitude_block_1_exp (None, 48, 48, 96)   385         prune_low_magnitude_block_1_expan
__________________________________________________________________________________________________
prune_low_magnitude_block_1_exp (None, 48, 48, 96)   1           prune_low_magnitude_block_1_expan
__________________________________________________________________________________________________
prune_low_magnitude_block_1_pad (None, 49, 49, 96)   1           prune_low_magnitude_block_1_expan
__________________________________________________________________________________________________
prune_low_magnitude_block_1_dep (None, 24, 24, 96)   865         prune_low_magnitude_block_1_pad[0
__________________________________________________________________________________________________
prune_low_magnitude_block_1_dep (None, 24, 24, 96)   385         prune_low_magnitude_block_1_depth
__________________________________________________________________________________________________
prune_low_magnitude_block_1_dep (None, 24, 24, 96)   1           prune_low_magnitude_block_1_depth
__________________________________________________________________________________________________
prune_low_magnitude_block_1_pro (None, 24, 24, 24)   4610        prune_low_magnitude_block_1_depth
__________________________________________________________________________________________________
prune_low_magnitude_block_1_pro (None, 24, 24, 24)   97          prune_low_magnitude_block_1_proje
__________________________________________________________________________________________________
prune_low_magnitude_block_2_exp (None, 24, 24, 144)  6914        prune_low_magnitude_block_1_proje
__________________________________________________________________________________________________
prune_low_magnitude_block_2_exp (None, 24, 24, 144)  577         prune_low_magnitude_block_2_expan
__________________________________________________________________________________________________
```



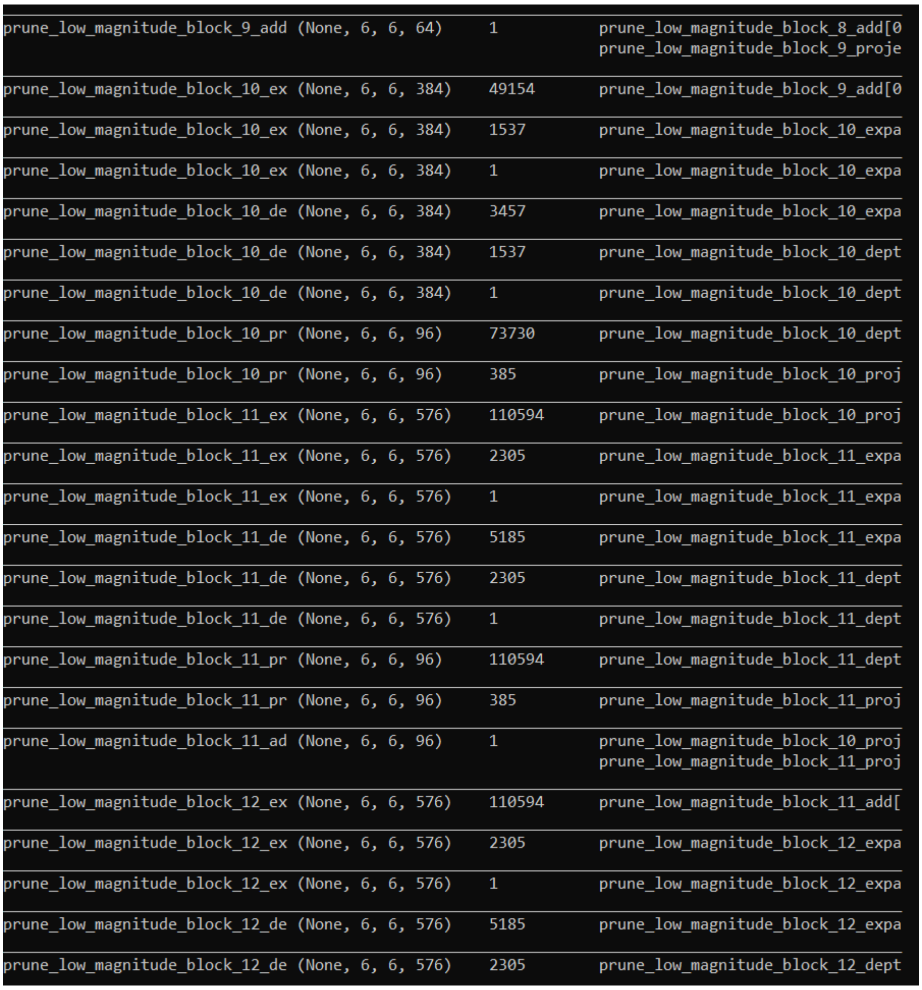
```
prune_low_magnitude_block_2_exp (None, 24, 24, 144)  1           prune_low_magnitude_block_2_expan

prune_low_magnitude_block_2_dep (None, 24, 24, 144)  1297        prune_low_magnitude_block_2_expan

prune_low_magnitude_block_2_dep (None, 24, 24, 144)  577         prune_low_magnitude_block_2_depth

prune_low_magnitude_block_2_dep (None, 24, 24, 144)  1           prune_low_magnitude_block_2_depth

prune_low_magnitude_block_2_pro (None, 24, 24, 24)   6914        prune_low_magnitude_block_2_depth

prune_low_magnitude_block_2_pro (None, 24, 24, 24)   97          prune_low_magnitude_block_2_proje

prune_low_magnitude_block_2_add (None, 24, 24, 24)   1           prune_low_magnitude_block_1_proje
                                                                 prune_low_magnitude_block_2_proje

prune_low_magnitude_block_3_exp (None, 24, 24, 144)  6914        prune_low_magnitude_block_2_add[0

prune_low_magnitude_block_3_exp (None, 24, 24, 144)  577         prune_low_magnitude_block_3_expan

prune_low_magnitude_block_3_exp (None, 24, 24, 144)  1           prune_low_magnitude_block_3_expan

prune_low_magnitude_block_3_pad (None, 25, 25, 144)  1           prune_low_magnitude_block_3_expan

prune_low_magnitude_block_3_dep (None, 12, 12, 144)  1297        prune_low_magnitude_block_3_pad[0

prune_low_magnitude_block_3_dep (None, 12, 12, 144)  577         prune_low_magnitude_block_3_depth

prune_low_magnitude_block_3_dep (None, 12, 12, 144)  1           prune_low_magnitude_block_3_depth

prune_low_magnitude_block_3_pro (None, 12, 12, 32)   9218        prune_low_magnitude_block_3_depth

prune_low_magnitude_block_3_pro (None, 12, 12, 32)   129         prune_low_magnitude_block_3_proje

prune_low_magnitude_block_4_exp (None, 12, 12, 192)  12290       prune_low_magnitude_block_3_proje

prune_low_magnitude_block_4_exp (None, 12, 12, 192)  769         prune_low_magnitude_block_4_expan

prune_low_magnitude_block_4_exp (None, 12, 12, 192)  1           prune_low_magnitude_block_4_expan

prune_low_magnitude_block_4_dep (None, 12, 12, 192)  1729        prune_low_magnitude_block_4_expan

prune_low_magnitude_block_4_dep (None, 12, 12, 192)  769         prune_low_magnitude_block_4_depth

prune_low_magnitude_block_4_dep (None, 12, 12, 192)  1           prune_low_magnitude_block_4_depth

prune_low_magnitude_block_4_pro (None, 12, 12, 32)   12290       prune_low_magnitude_block_4_depth
```



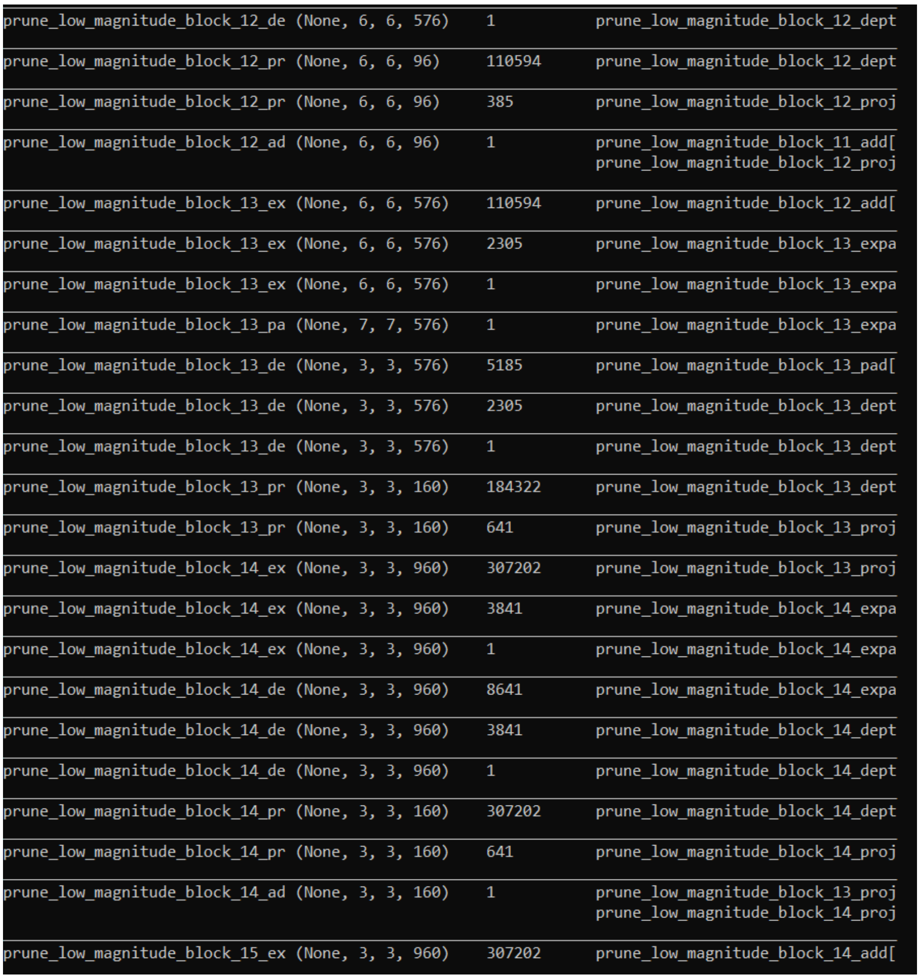
```
prune_low_magnitude_block_4_pro (None, 12, 12, 32)   129         prune_low_magnitude_block_4_proje

prune_low_magnitude_block_4_add (None, 12, 12, 32)   1           prune_low_magnitude_block_3_proje
                                                                 prune_low_magnitude_block_4_proje

prune_low_magnitude_block_5_exp (None, 12, 12, 192)  12290       prune_low_magnitude_block_4_add[0

prune_low_magnitude_block_5_exp (None, 12, 12, 192)  769         prune_low_magnitude_block_5_expan

prune_low_magnitude_block_5_exp (None, 12, 12, 192)  1           prune_low_magnitude_block_5_expan

prune_low_magnitude_block_5_dep (None, 12, 12, 192)  1729        prune_low_magnitude_block_5_expan

prune_low_magnitude_block_5_dep (None, 12, 12, 192)  769         prune_low_magnitude_block_5_depth

prune_low_magnitude_block_5_dep (None, 12, 12, 192)  1           prune_low_magnitude_block_5_depth

prune_low_magnitude_block_5_pro (None, 12, 12, 32)   12290       prune_low_magnitude_block_5_depth

prune_low_magnitude_block_5_pro (None, 12, 12, 32)   129         prune_low_magnitude_block_5_proje

prune_low_magnitude_block_5_add (None, 12, 12, 32)   1           prune_low_magnitude_block_4_add[0
                                                                 prune_low_magnitude_block_5_proje

prune_low_magnitude_block_6_exp (None, 12, 12, 192)  12290       prune_low_magnitude_block_5_add[0

prune_low_magnitude_block_6_exp (None, 12, 12, 192)  769         prune_low_magnitude_block_6_expan

prune_low_magnitude_block_6_exp (None, 12, 12, 192)  1           prune_low_magnitude_block_6_expan

prune_low_magnitude_block_6_pad (None, 13, 13, 192)  1           prune_low_magnitude_block_6_expan

prune_low_magnitude_block_6_dep (None, 6, 6, 192)    1729        prune_low_magnitude_block_6_pad[0

prune_low_magnitude_block_6_dep (None, 6, 6, 192)    769         prune_low_magnitude_block_6_depth

prune_low_magnitude_block_6_dep (None, 6, 6, 192)    1           prune_low_magnitude_block_6_depth

prune_low_magnitude_block_6_pro (None, 6, 6, 64)     24578       prune_low_magnitude_block_6_depth

prune_low_magnitude_block_6_pro (None, 6, 6, 64)     257         prune_low_magnitude_block_6_proje

prune_low_magnitude_block_7_exp (None, 6, 6, 384)    49154       prune_low_magnitude_block_6_proje

prune_low_magnitude_block_7_exp (None, 6, 6, 384)    1537        prune_low_magnitude_block_7_expan

prune_low_magnitude_block_7_exp (None, 6, 6, 384)    1           prune_low_magnitude_block_7_expan
```



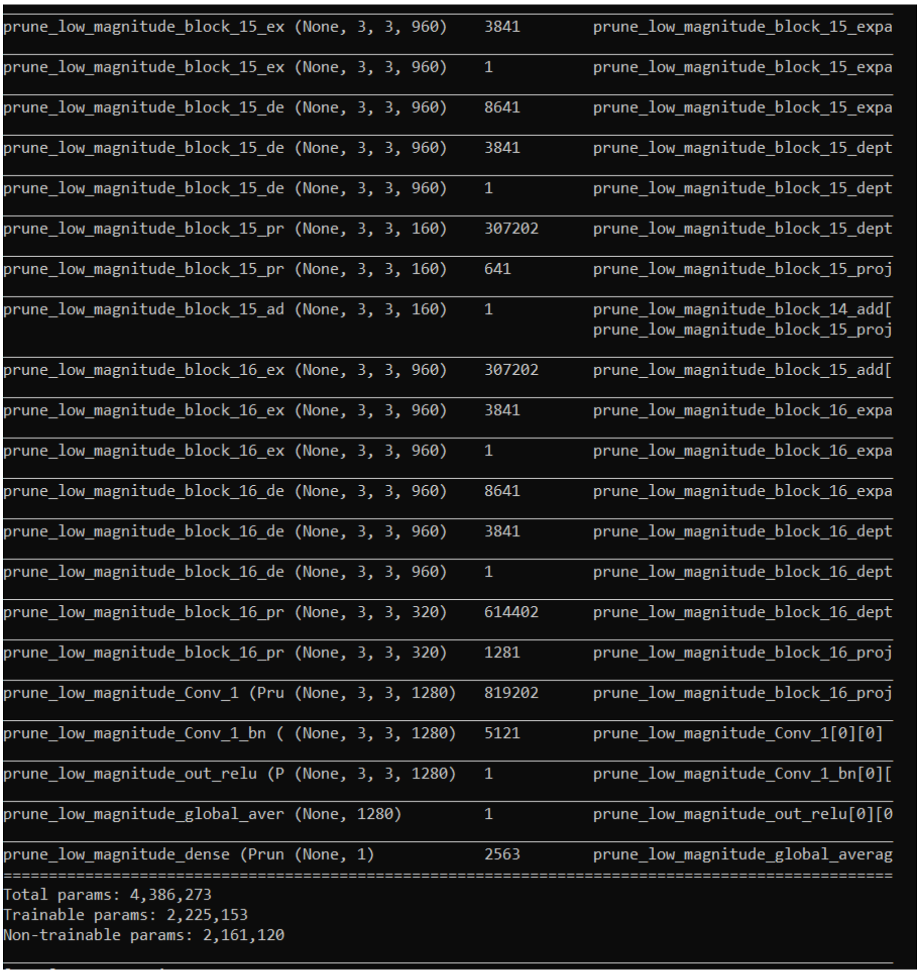
```
prune_low_magnitude_block_7_dep (None, 6, 6, 384)    3457        prune_low_magnitude_block_7_expan

prune_low_magnitude_block_7_dep (None, 6, 6, 384)    1537        prune_low_magnitude_block_7_depth

prune_low_magnitude_block_7_dep (None, 6, 6, 384)    1           prune_low_magnitude_block_7_depth

prune_low_magnitude_block_7_pro (None, 6, 6, 64)     49154       prune_low_magnitude_block_7_depth

prune_low_magnitude_block_7_pro (None, 6, 6, 64)     257         prune_low_magnitude_block_7_proje

prune_low_magnitude_block_7_add (None, 6, 6, 64)     1           prune_low_magnitude_block_6_proje
                                                                 prune_low_magnitude_block_7_proje

prune_low_magnitude_block_8_exp (None, 6, 6, 384)    49154       prune_low_magnitude_block_7_add[0

prune_low_magnitude_block_8_exp (None, 6, 6, 384)    1537        prune_low_magnitude_block_8_expan

prune_low_magnitude_block_8_exp (None, 6, 6, 384)    1           prune_low_magnitude_block_8_expan

prune_low_magnitude_block_8_dep (None, 6, 6, 384)    3457        prune_low_magnitude_block_8_expan

prune_low_magnitude_block_8_dep (None, 6, 6, 384)    1537        prune_low_magnitude_block_8_depth

prune_low_magnitude_block_8_dep (None, 6, 6, 384)    1           prune_low_magnitude_block_8_depth

prune_low_magnitude_block_8_pro (None, 6, 6, 64)     49154       prune_low_magnitude_block_8_depth

prune_low_magnitude_block_8_pro (None, 6, 6, 64)     257         prune_low_magnitude_block_8_proje

prune_low_magnitude_block_8_add (None, 6, 6, 64)     1           prune_low_magnitude_block_7_add[0
                                                                 prune_low_magnitude_block_8_proje

prune_low_magnitude_block_9_exp (None, 6, 6, 384)    49154       prune_low_magnitude_block_8_add[0

prune_low_magnitude_block_9_exp (None, 6, 6, 384)    1537        prune_low_magnitude_block_9_expan

prune_low_magnitude_block_9_exp (None, 6, 6, 384)    1           prune_low_magnitude_block_9_expan

prune_low_magnitude_block_9_dep (None, 6, 6, 384)    3457        prune_low_magnitude_block_9_expan

prune_low_magnitude_block_9_dep (None, 6, 6, 384)    1537        prune_low_magnitude_block_9_depth

prune_low_magnitude_block_9_dep (None, 6, 6, 384)    1           prune_low_magnitude_block_9_depth

prune_low_magnitude_block_9_pro (None, 6, 6, 64)     49154       prune_low_magnitude_block_9_depth

prune_low_magnitude_block_9_pro (None, 6, 6, 64)     257         prune_low_magnitude_block_9_proje
```



```
prune_low_magnitude_block_9_add (None, 6, 6, 64)     1            prune_low_magnitude_block_8_add[0
                                                                  prune_low_magnitude_block_9_proje

prune_low_magnitude_block_10_ex (None, 6, 6, 384)    49154        prune_low_magnitude_block_9_add[0

prune_low_magnitude_block_10_ex (None, 6, 6, 384)    1537         prune_low_magnitude_block_10_expa

prune_low_magnitude_block_10_ex (None, 6, 6, 384)    1            prune_low_magnitude_block_10_expa

prune_low_magnitude_block_10_de (None, 6, 6, 384)    3457         prune_low_magnitude_block_10_expa

prune_low_magnitude_block_10_de (None, 6, 6, 384)    1537         prune_low_magnitude_block_10_dept

prune_low_magnitude_block_10_de (None, 6, 6, 384)    1            prune_low_magnitude_block_10_dept

prune_low_magnitude_block_10_pr (None, 6, 6, 96)     73730        prune_low_magnitude_block_10_dept

prune_low_magnitude_block_10_pr (None, 6, 6, 96)     385          prune_low_magnitude_block_10_proj

prune_low_magnitude_block_11_ex (None, 6, 6, 576)    110594       prune_low_magnitude_block_10_proj

prune_low_magnitude_block_11_ex (None, 6, 6, 576)    2305         prune_low_magnitude_block_11_expa

prune_low_magnitude_block_11_ex (None, 6, 6, 576)    1            prune_low_magnitude_block_11_expa

prune_low_magnitude_block_11_de (None, 6, 6, 576)    5185         prune_low_magnitude_block_11_expa

prune_low_magnitude_block_11_de (None, 6, 6, 576)    2305         prune_low_magnitude_block_11_dept

prune_low_magnitude_block_11_de (None, 6, 6, 576)    1            prune_low_magnitude_block_11_dept

prune_low_magnitude_block_11_pr (None, 6, 6, 96)     110594       prune_low_magnitude_block_11_dept

prune_low_magnitude_block_11_pr (None, 6, 6, 96)     385          prune_low_magnitude_block_11_proj

prune_low_magnitude_block_11_ad (None, 6, 6, 96)     1            prune_low_magnitude_block_10_proj
                                                                  prune_low_magnitude_block_11_proj

prune_low_magnitude_block_12_ex (None, 6, 6, 576)    110594       prune_low_magnitude_block_11_add[

prune_low_magnitude_block_12_ex (None, 6, 6, 576)    2305         prune_low_magnitude_block_12_expa

prune_low_magnitude_block_12_ex (None, 6, 6, 576)    1            prune_low_magnitude_block_12_expa

prune_low_magnitude_block_12_de (None, 6, 6, 576)    5185         prune_low_magnitude_block_12_expa

prune_low_magnitude_block_12_de (None, 6, 6, 576)    2305         prune_low_magnitude_block_12_dept
```

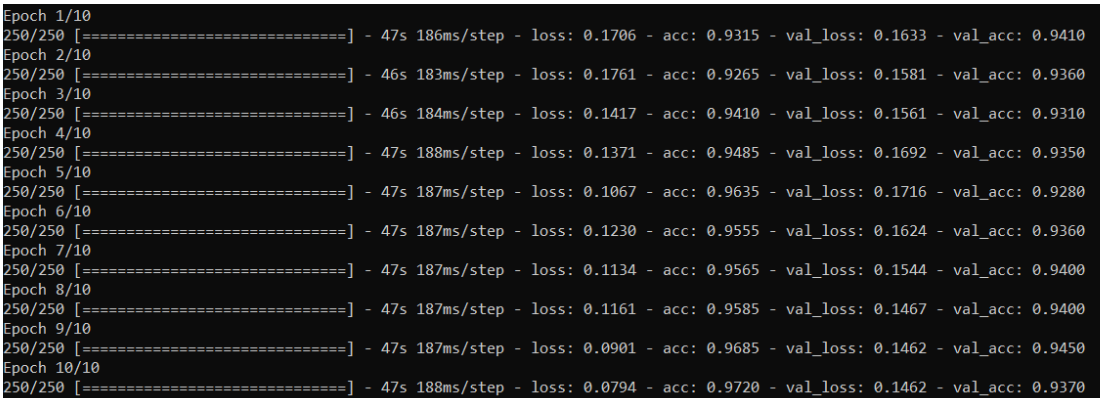


```
prune_low_magnitude_block_12_de (None, 6, 6, 576)    1          prune_low_magnitude_block_12_dept

prune_low_magnitude_block_12_pr (None, 6, 6, 96)     110594     prune_low_magnitude_block_12_dept

prune_low_magnitude_block_12_pr (None, 6, 6, 96)     385        prune_low_magnitude_block_12_proj

prune_low_magnitude_block_12_ad (None, 6, 6, 96)     1          prune_low_magnitude_block_11_add[
                                                                prune_low_magnitude_block_12_proj

prune_low_magnitude_block_13_ex (None, 6, 6, 576)    110594     prune_low_magnitude_block_12_add[

prune_low_magnitude_block_13_ex (None, 6, 6, 576)    2305       prune_low_magnitude_block_13_expa

prune_low_magnitude_block_13_ex (None, 6, 6, 576)    1          prune_low_magnitude_block_13_expa

prune_low_magnitude_block_13_pa (None, 7, 7, 576)    1          prune_low_magnitude_block_13_expa

prune_low_magnitude_block_13_de (None, 3, 3, 576)    5185       prune_low_magnitude_block_13_pad[

prune_low_magnitude_block_13_de (None, 3, 3, 576)    2305       prune_low_magnitude_block_13_dept

prune_low_magnitude_block_13_de (None, 3, 3, 576)    1          prune_low_magnitude_block_13_dept

prune_low_magnitude_block_13_pr (None, 3, 3, 160)    184322     prune_low_magnitude_block_13_dept

prune_low_magnitude_block_13_pr (None, 3, 3, 160)    641        prune_low_magnitude_block_13_proj

prune_low_magnitude_block_14_ex (None, 3, 3, 960)    307202     prune_low_magnitude_block_13_proj

prune_low_magnitude_block_14_ex (None, 3, 3, 960)    3841       prune_low_magnitude_block_14_expa

prune_low_magnitude_block_14_ex (None, 3, 3, 960)    1          prune_low_magnitude_block_14_expa

prune_low_magnitude_block_14_de (None, 3, 3, 960)    8641       prune_low_magnitude_block_14_expa

prune_low_magnitude_block_14_de (None, 3, 3, 960)    3841       prune_low_magnitude_block_14_dept

prune_low_magnitude_block_14_de (None, 3, 3, 960)    1          prune_low_magnitude_block_14_dept

prune_low_magnitude_block_14_pr (None, 3, 3, 160)    307202     prune_low_magnitude_block_14_dept

prune_low_magnitude_block_14_pr (None, 3, 3, 160)    641        prune_low_magnitude_block_14_proj

prune_low_magnitude_block_14_ad (None, 3, 3, 160)    1          prune_low_magnitude_block_13_proj
                                                                prune_low_magnitude_block_14_proj

prune_low_magnitude_block_15_ex (None, 3, 3, 960)    307202     prune_low_magnitude_block_14_add[
```



```
prune_low_magnitude_block_15_ex (None, 3, 3, 960)    3841        prune_low_magnitude_block_15_expa
__________________________________________________________________________________________________
prune_low_magnitude_block_15_ex (None, 3, 3, 960)    1           prune_low_magnitude_block_15_expa
__________________________________________________________________________________________________
prune_low_magnitude_block_15_de (None, 3, 3, 960)    8641        prune_low_magnitude_block_15_expa
__________________________________________________________________________________________________
prune_low_magnitude_block_15_de (None, 3, 3, 960)    3841        prune_low_magnitude_block_15_dept
__________________________________________________________________________________________________
prune_low_magnitude_block_15_de (None, 3, 3, 960)    1           prune_low_magnitude_block_15_dept
__________________________________________________________________________________________________
prune_low_magnitude_block_15_pr (None, 3, 3, 160)    307202      prune_low_magnitude_block_15_dept
__________________________________________________________________________________________________
prune_low_magnitude_block_15_pr (None, 3, 3, 160)    641         prune_low_magnitude_block_15_proj
__________________________________________________________________________________________________
prune_low_magnitude_block_15_ad (None, 3, 3, 160)    1           prune_low_magnitude_block_14_add[
                                                                 prune_low_magnitude_block_15_proj
__________________________________________________________________________________________________
prune_low_magnitude_block_16_ex (None, 3, 3, 960)    307202      prune_low_magnitude_block_15_add[
__________________________________________________________________________________________________
prune_low_magnitude_block_16_ex (None, 3, 3, 960)    3841        prune_low_magnitude_block_16_expa
__________________________________________________________________________________________________
prune_low_magnitude_block_16_ex (None, 3, 3, 960)    1           prune_low_magnitude_block_16_expa
__________________________________________________________________________________________________
prune_low_magnitude_block_16_de (None, 3, 3, 960)    8641        prune_low_magnitude_block_16_expa
__________________________________________________________________________________________________
prune_low_magnitude_block_16_de (None, 3, 3, 960)    3841        prune_low_magnitude_block_16_dept
__________________________________________________________________________________________________
prune_low_magnitude_block_16_de (None, 3, 3, 960)    1           prune_low_magnitude_block_16_dept
__________________________________________________________________________________________________
prune_low_magnitude_block_16_pr (None, 3, 3, 320)    614402      prune_low_magnitude_block_16_dept
__________________________________________________________________________________________________
prune_low_magnitude_block_16_pr (None, 3, 3, 320)    1281        prune_low_magnitude_block_16_proj
__________________________________________________________________________________________________
prune_low_magnitude_Conv_1 (Pru (None, 3, 3, 1280)   819202      prune_low_magnitude_block_16_proj
__________________________________________________________________________________________________
prune_low_magnitude_Conv_1_bn ( (None, 3, 3, 1280)   5121        prune_low_magnitude_Conv_1[0][0]
__________________________________________________________________________________________________
prune_low_magnitude_out_relu (P (None, 3, 3, 1280)   1           prune_low_magnitude_Conv_1_bn[0][
__________________________________________________________________________________________________
prune_low_magnitude_global_aver (None, 1280)         1           prune_low_magnitude_out_relu[0][0
__________________________________________________________________________________________________
prune_low_magnitude_dense (Prun (None, 1)            2563        prune_low_magnitude_global_averag
==================================================================================================
Total params: 4,386,273
Trainable params: 2,225,153
Non-trainable params: 2,161,120
__________________________________________________________________________________________________
```

1.2 Pruning log



```
Epoch 1/10
250/250 [==============================] - 47s 186ms/step - loss: 0.1706 - acc: 0.9315 - val_loss: 0.1633 - val_acc: 0.9410
Epoch 2/10
250/250 [==============================] - 46s 183ms/step - loss: 0.1761 - acc: 0.9265 - val_loss: 0.1581 - val_acc: 0.9360
Epoch 3/10
250/250 [==============================] - 46s 184ms/step - loss: 0.1417 - acc: 0.9410 - val_loss: 0.1561 - val_acc: 0.9310
Epoch 4/10
250/250 [==============================] - 47s 188ms/step - loss: 0.1371 - acc: 0.9485 - val_loss: 0.1692 - val_acc: 0.9350
Epoch 5/10
250/250 [==============================] - 47s 187ms/step - loss: 0.1067 - acc: 0.9635 - val_loss: 0.1716 - val_acc: 0.9280
Epoch 6/10
250/250 [==============================] - 47s 187ms/step - loss: 0.1230 - acc: 0.9555 - val_loss: 0.1624 - val_acc: 0.9360
Epoch 7/10
250/250 [==============================] - 47s 187ms/step - loss: 0.1134 - acc: 0.9565 - val_loss: 0.1544 - val_acc: 0.9400
Epoch 8/10
250/250 [==============================] - 47s 187ms/step - loss: 0.1161 - acc: 0.9585 - val_loss: 0.1467 - val_acc: 0.9400
Epoch 9/10
250/250 [==============================] - 47s 187ms/step - loss: 0.0901 - acc: 0.9685 - val_loss: 0.1462 - val_acc: 0.9450
Epoch 10/10
250/250 [==============================] - 47s 188ms/step - loss: 0.0794 - acc: 0.9720 - val_loss: 0.1462 - val_acc: 0.9370
```

1.3 Comparison

Final sparsity	0.3 (original)	0.25	0.5	0.75
Compressed model size (MB)	6.43	6.73	5.11	3.29
Validation acc	94.50%	94.10%	90.20%	63.80%

1.4 Description of the steps of the pruning process

Step 1: Start the training job

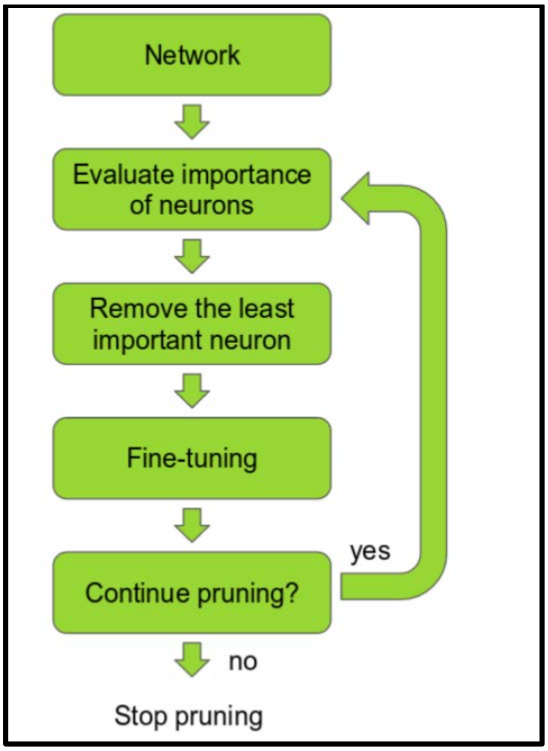
Step 2: Acquire the weights, gradients, biases, and activation outputs

Step 3: Compute filter ranks

Step 4: Prune low-ranking filters

Step 5: Set new weights

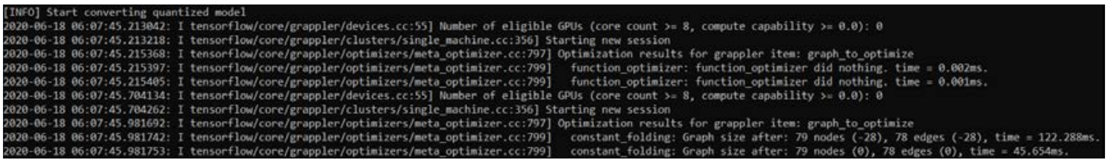
Step 6: Start the training job with the pruned model


Network
Evaluate importance of neurons
Remove the least important neuron
Fine-tuning
Continue pruning?
yes
no
Stop pruning

taken from https://jacobgil.github.io/deeplearning/pruning-deep-learning

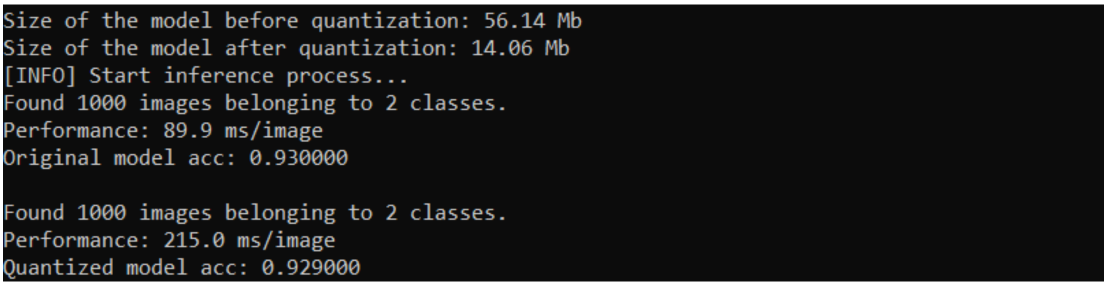
2. Task 2

2.1 Quantization process



```
[INFO] Start converting quantized model
2020-06-18 06:07:45.213042: I tensorflow/core/grappler/devices.cc:55] Number of eligible GPUs (core count >= 8, compute capability >= 0.0): 0
2020-06-18 06:07:45.213218: I tensorflow/core/grappler/clusters/single_machine.cc:356] Starting new session
2020-06-18 06:07:45.215368: I tensorflow/core/grappler/optimizers/meta_optimizer.cc:797] Optimization results for grappler item: graph_to_optimize
2020-06-18 06:07:45.215397: I tensorflow/core/grappler/optimizers/meta_optimizer.cc:799]   function_optimizer: function_optimizer did nothing. time = 0.002ms.
2020-06-18 06:07:45.215405: I tensorflow/core/grappler/optimizers/meta_optimizer.cc:799]   function_optimizer: function_optimizer did nothing. time = 0.001ms.
2020-06-18 06:07:45.704134: I tensorflow/core/grappler/devices.cc:55] Number of eligible GPUs (core count >= 8, compute capability >= 0.0): 0
2020-06-18 06:07:45.704262: I tensorflow/core/grappler/clusters/single_machine.cc:356] Starting new session
2020-06-18 06:07:45.981692: I tensorflow/core/grappler/optimizers/meta_optimizer.cc:797] Optimization results for grappler item: graph_to_optimize
2020-06-18 06:07:45.981742: I tensorflow/core/grappler/optimizers/meta_optimizer.cc:799]   constant_folding: Graph size after: 79 nodes (-28), 78 edges (-28), time = 122.288ms.
2020-06-18 06:07:45.981753: I tensorflow/core/grappler/optimizers/meta_optimizer.cc:799]   constant_folding: Graph size after: 79 nodes (0), 78 edges (0), time = 45.654ms.
```

2.2 Validation accuracy



```
Size of the model before quantization: 56.14 Mb
Size of the model after quantization: 14.06 Mb
[INFO] Start inference process...
Found 1000 images belonging to 2 classes.
Performance: 89.9 ms/image
Original model acc: 0.930000

Found 1000 images belonging to 2 classes.
Performance: 215.0 ms/image
Quantized model acc: 0.929000
```

2.3 Comparison

Quantization	None (Original)	Dynamic Range	FP16
tflite file’s size (MB)	56.14	14.06	28.08
Validation acc	93.00%	92.90%	93.00%

2.4 Description of the difference between post-training quantization and quantization aware training

There are two forms of quantization:

a. Post-training quantization
   - Easier to use
   - Includes techniques to reduce CPU and hardware accelerator latency, processing power, and model size with little degradation in model accuracy.

b. Quantization aware training
   - Often better for model accuracy
   - Emulates inference-time quantization, creating a model that downstream tools will use to produce the quantized model.
   - The quantized models use lower-precision, such as 8-bit, leading to benefits during deployment.

3. Advance

3.1 Modifications

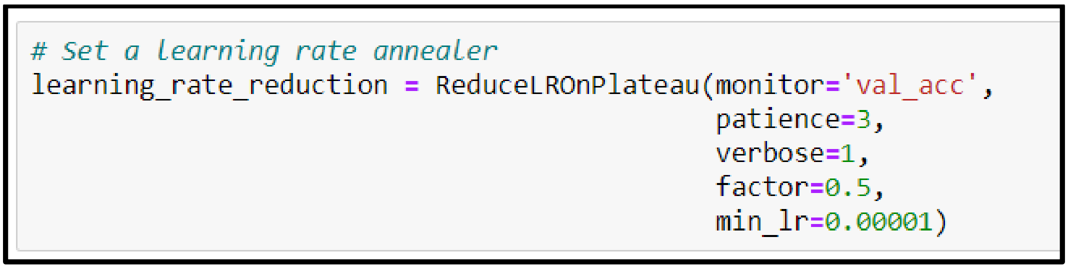
- The value of epochs

  I changed the value of epochs to 15 for both training on VGG16 model and pruned VGG16 model
- The width and height of images

  I modified the size of images from (112, 112) to (150, 150)
- Final sparsity

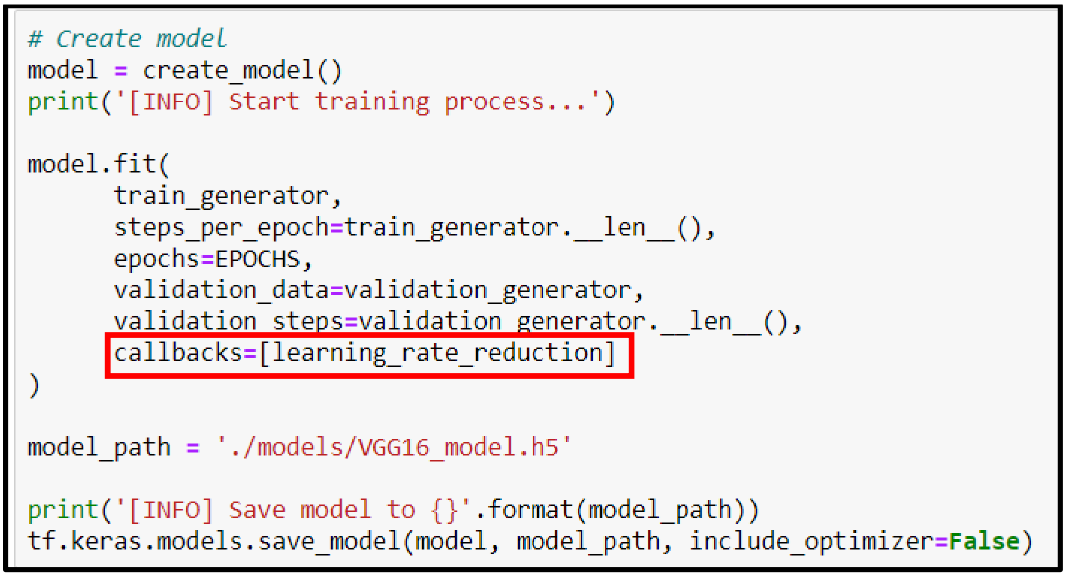
  As stated in hw3.pdf, the final sparsity must be set as 0.9
- Callbacks

  I have added a learning reduction to callbacks during the model training



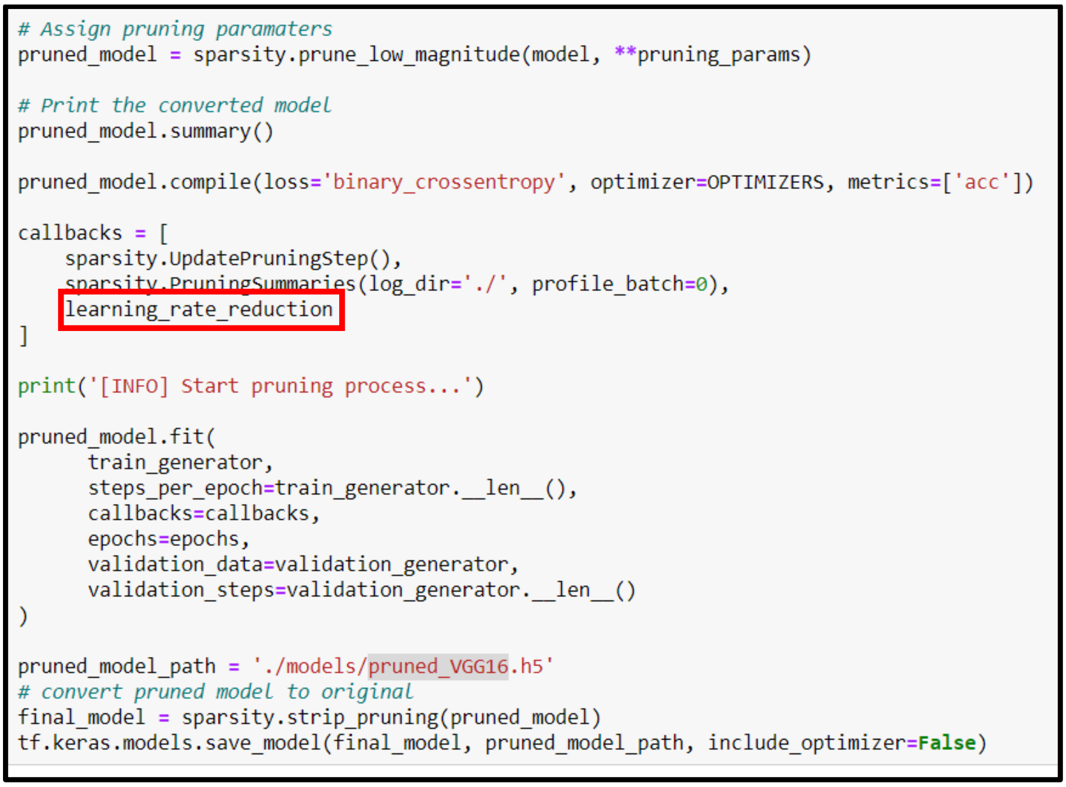
```
# Set a learning rate annealer
learning_rate_reduction = ReduceLROnPlateau(monitor='val_acc',
                                            patience=3,
                                            verbose=1,
                                            factor=0.5,
                                            min_lr=0.00001)
```

During VGG16 training



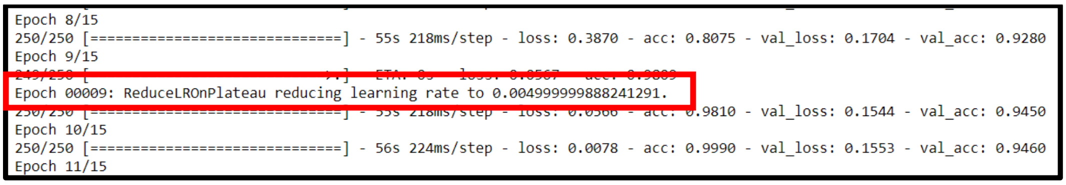
```
# Create model
model = create_model()
print('[INFO] Start training process...')

model.fit(
      train_generator,
      steps_per_epoch=train_generator.__len__(),
      epochs=EPOCHS,
      validation_data=validation_generator,
      validation_steps=validation_generator.__len__(),
      callbacks=[learning_rate_reduction]
)

model_path = './models/VGG16_model.h5'

print('[INFO] Save model to {}'.format(model_path))
tf.keras.models.save_model(model, model_path, include_optimizer=False)
```

During pruned_VGG16 training



```
# Assign pruning paramaters
pruned_model = sparsity.prune_low_magnitude(model, **pruning_params)

# Print the converted model
pruned_model.summary()

pruned_model.compile(loss='binary_crossentropy', optimizer=OPTIMIZERS, metrics=['acc'])

callbacks = [
    sparsity.UpdatePruningStep(),
    sparsity.PruningSummaries(log_dir='./', profile_batch=0),
    learning_rate_reduction
]

print('[INFO] Start pruning process...')

pruned_model.fit(
      train_generator,
      steps_per_epoch=train_generator.__len__(),
      callbacks=callbacks,
      epochs=epochs,
      validation_data=validation_generator,
      validation_steps=validation_generator.__len__()
)

pruned_model_path = './models/pruned_VGG16.h5'
# convert pruned model to original
final_model = sparsity.strip_pruning(pruned_model)
tf.keras.models.save_model(final_model, pruned_model_path, include_optimizer=False)
```

Learning Reduction during the training process



```
Epoch 8/15
250/250 [==============================] - 55s 218ms/step - loss: 0.3870 - acc: 0.8075 - val_loss: 0.1704 - val_acc: 0.9280
Epoch 9/15
249/250 [============================>.] - ETA: 0s - loss: 0.0567 - acc: 0.9809
Epoch 00009: ReduceLROnPlateau reducing learning rate to 0.004999999888241291.
250/250 [==============================] - 55s 218ms/step - loss: 0.0566 - acc: 0.9810 - val_loss: 0.1544 - val_acc: 0.9450
Epoch 10/15
250/250 [==============================] - 56s 224ms/step - loss: 0.0078 - acc: 0.9990 - val_loss: 0.1553 - val_acc: 0.9460
Epoch 11/15
```

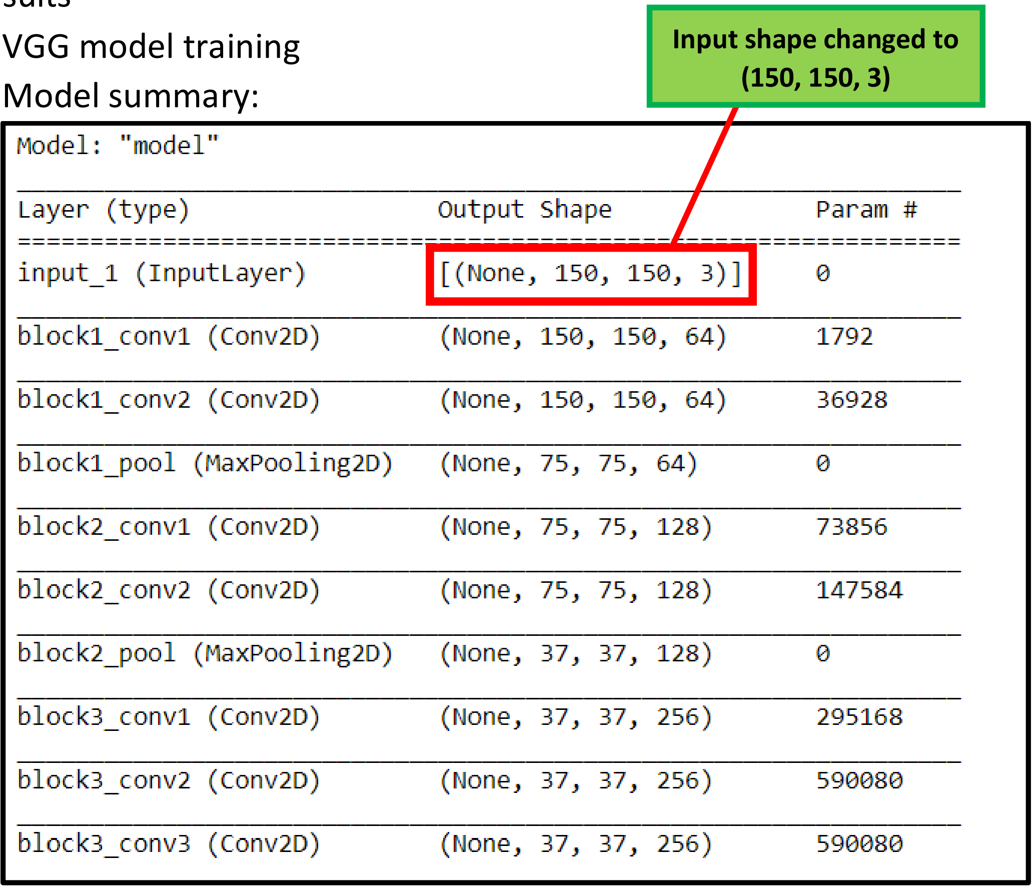
- Other settings I have tried but haven't added them

  I have tried to modify the optimizers such as Adam, Adagrad, Adamax, Nadam, Ftrl, RMSprop, and so on, but the result comes out the SGD would be the best optimizer on this case.

3.2 Results

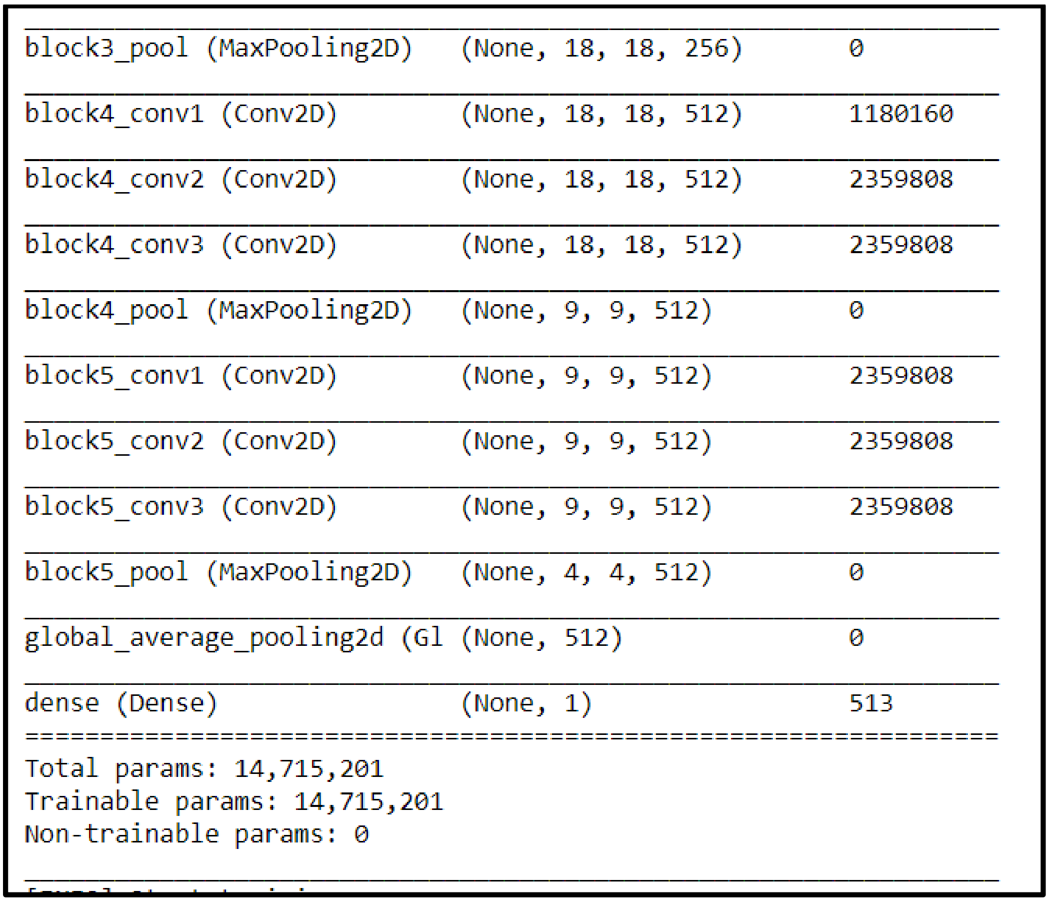
- VGG model training

  Model summary:



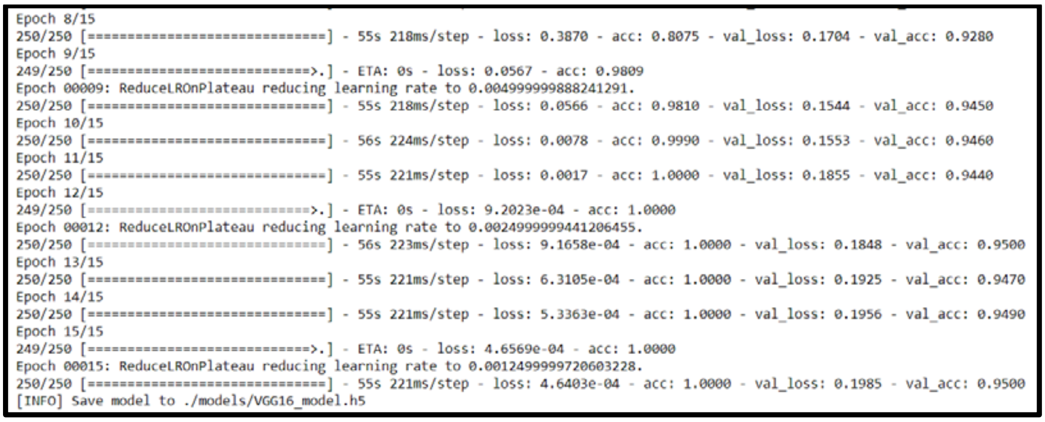
Input shape changed to (150, 150, 3)

```
Model: "model"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         [(None, 150, 150, 3)]     0
_________________________________________________________________
block1_conv1 (Conv2D)        (None, 150, 150, 64)      1792
_________________________________________________________________
block1_conv2 (Conv2D)        (None, 150, 150, 64)      36928
_________________________________________________________________
block1_pool (MaxPooling2D)   (None, 75, 75, 64)        0
_________________________________________________________________
block2_conv1 (Conv2D)        (None, 75, 75, 128)       73856
_________________________________________________________________
block2_conv2 (Conv2D)        (None, 75, 75, 128)       147584
_________________________________________________________________
block2_pool (MaxPooling2D)   (None, 37, 37, 128)       0
_________________________________________________________________
block3_conv1 (Conv2D)        (None, 37, 37, 256)       295168
_________________________________________________________________
block3_conv2 (Conv2D)        (None, 37, 37, 256)       590080
_________________________________________________________________
block3_conv3 (Conv2D)        (None, 37, 37, 256)       590080
```



```
_________________________________________________________________
block3_pool (MaxPooling2D)   (None, 18, 18, 256)       0
_________________________________________________________________
block4_conv1 (Conv2D)        (None, 18, 18, 512)       1180160
_________________________________________________________________
block4_conv2 (Conv2D)        (None, 18, 18, 512)       2359808
_________________________________________________________________
block4_conv3 (Conv2D)        (None, 18, 18, 512)       2359808
_________________________________________________________________
block4_pool (MaxPooling2D)   (None, 9, 9, 512)         0
_________________________________________________________________
block5_conv1 (Conv2D)        (None, 9, 9, 512)         2359808
_________________________________________________________________
block5_conv2 (Conv2D)        (None, 9, 9, 512)         2359808
_________________________________________________________________
block5_conv3 (Conv2D)        (None, 9, 9, 512)         2359808
_________________________________________________________________
block5_pool (MaxPooling2D)   (None, 4, 4, 512)         0
_________________________________________________________________
global_average_pooling2d (Gl (None, 512)               0
_________________________________________________________________
dense (Dense)                (None, 1)                 513
=================================================================
Total params: 14,715,201
Trainable params: 14,715,201
Non-trainable params: 0

_________________________________________________________________
```

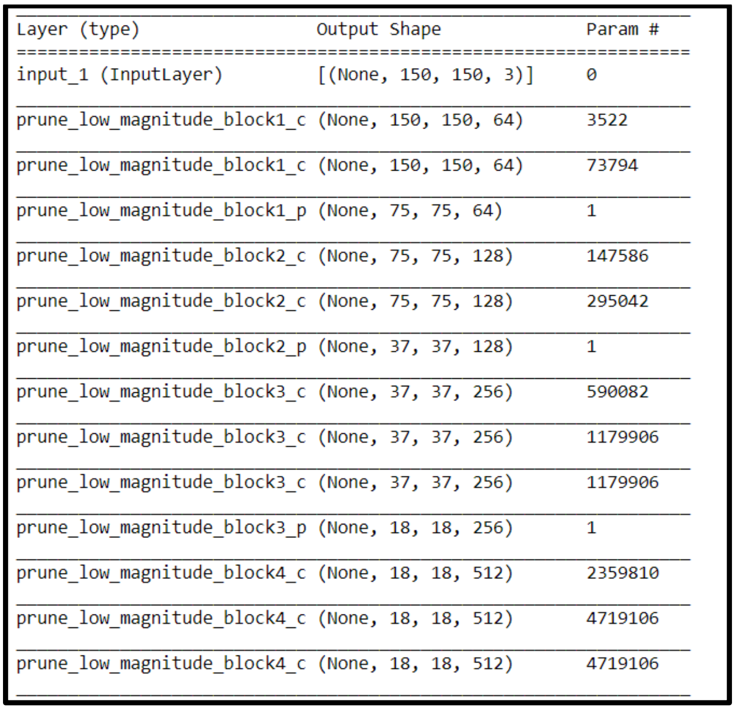
Training



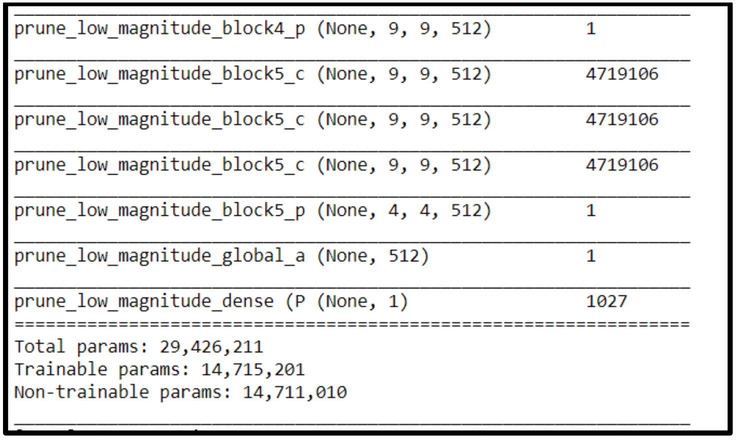
```
Epoch 8/15
250/250 [==============================] - 55s 218ms/step - loss: 0.3870 - acc: 0.8075 - val_loss: 0.1704 - val_acc: 0.9280
Epoch 9/15
249/250 [============================>.] - ETA: 0s - loss: 0.0567 - acc: 0.9809
Epoch 00009: ReduceLROnPlateau reducing learning rate to 0.004999999888241291.
250/250 [==============================] - 55s 218ms/step - loss: 0.0566 - acc: 0.9810 - val_loss: 0.1544 - val_acc: 0.9450
Epoch 10/15
250/250 [==============================] - 56s 224ms/step - loss: 0.0078 - acc: 0.9990 - val_loss: 0.1553 - val_acc: 0.9460
Epoch 11/15
250/250 [==============================] - 55s 221ms/step - loss: 0.0017 - acc: 1.0000 - val_loss: 0.1855 - val_acc: 0.9440
Epoch 12/15
249/250 [============================>.] - ETA: 0s - loss: 9.2023e-04 - acc: 1.0000
Epoch 00012: ReduceLROnPlateau reducing learning rate to 0.002499999441206455.
250/250 [==============================] - 56s 223ms/step - loss: 9.1658e-04 - acc: 1.0000 - val_loss: 0.1848 - val_acc: 0.9500
Epoch 13/15
250/250 [==============================] - 55s 221ms/step - loss: 6.3105e-04 - acc: 1.0000 - val_loss: 0.1925 - val_acc: 0.9470
Epoch 14/15
250/250 [==============================] - 55s 221ms/step - loss: 5.3363e-04 - acc: 1.0000 - val_loss: 0.1956 - val_acc: 0.9490
Epoch 15/15
249/250 [============================>.] - ETA: 0s - loss: 4.6569e-04 - acc: 1.0000
Epoch 00015: ReduceLROnPlateau reducing learning rate to 0.0012499999720603228.
250/250 [==============================] - 55s 221ms/step - loss: 4.6403e-04 - acc: 1.0000 - val_loss: 0.1985 - val_acc: 0.9500
[INFO] Save model to ./models/VGG16_model.h5
```

- Pruning

  Model summary



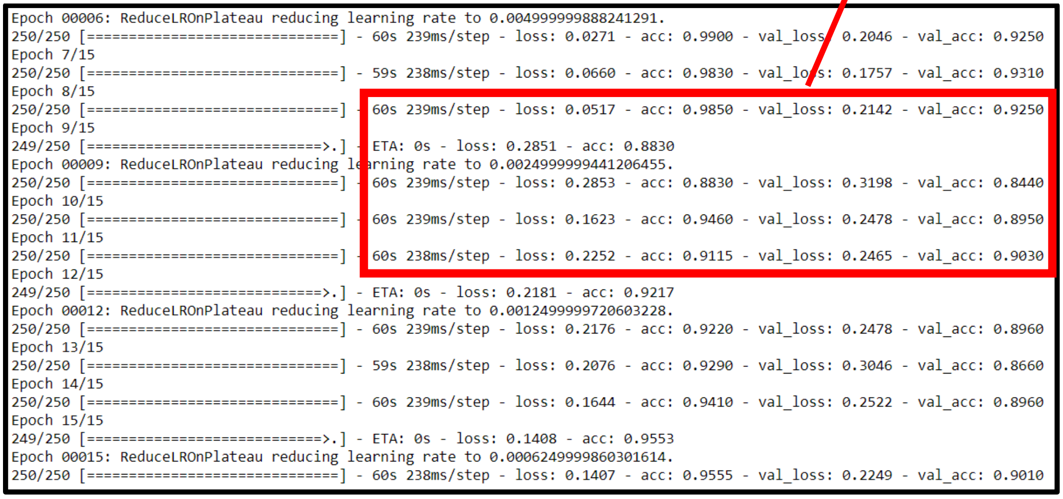
```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
input_1 (InputLayer)         [(None, 150, 150, 3)]     0         
_________________________________________________________________
prune_low_magnitude_block1_c (None, 150, 150, 64)      3522      
_________________________________________________________________
prune_low_magnitude_block1_c (None, 150, 150, 64)      73794     
_________________________________________________________________
prune_low_magnitude_block1_p (None, 75, 75, 64)        1         
_________________________________________________________________
prune_low_magnitude_block2_c (None, 75, 75, 128)       147586    
_________________________________________________________________
prune_low_magnitude_block2_c (None, 75, 75, 128)       295042    
_________________________________________________________________
prune_low_magnitude_block2_p (None, 37, 37, 128)       1         
_________________________________________________________________
prune_low_magnitude_block3_c (None, 37, 37, 256)       590082    
_________________________________________________________________
prune_low_magnitude_block3_c (None, 37, 37, 256)       1179906   
_________________________________________________________________
prune_low_magnitude_block3_c (None, 37, 37, 256)       1179906   
_________________________________________________________________
prune_low_magnitude_block3_p (None, 18, 18, 256)       1         
_________________________________________________________________
prune_low_magnitude_block4_c (None, 18, 18, 512)       2359810   
_________________________________________________________________
prune_low_magnitude_block4_c (None, 18, 18, 512)       4719106   
_________________________________________________________________
prune_low_magnitude_block4_c (None, 18, 18, 512)       4719106   
_________________________________________________________________
```



```
_________________________________________________________________
prune_low_magnitude_block4_p (None, 9, 9, 512)         1         
_________________________________________________________________
prune_low_magnitude_block5_c (None, 9, 9, 512)         4719106   
_________________________________________________________________
prune_low_magnitude_block5_c (None, 9, 9, 512)         4719106   
_________________________________________________________________
prune_low_magnitude_block5_c (None, 9, 9, 512)         4719106   
_________________________________________________________________
prune_low_magnitude_block5_p (None, 4, 4, 512)         1         
_________________________________________________________________
prune_low_magnitude_global_a (None, 512)               1         
_________________________________________________________________
prune_low_magnitude_dense (P (None, 1)                 1027      
=================================================================
Total params: 29,426,211
Trainable params: 14,715,201
Non-trainable params: 14,711,010
_________________________________________________________________
```

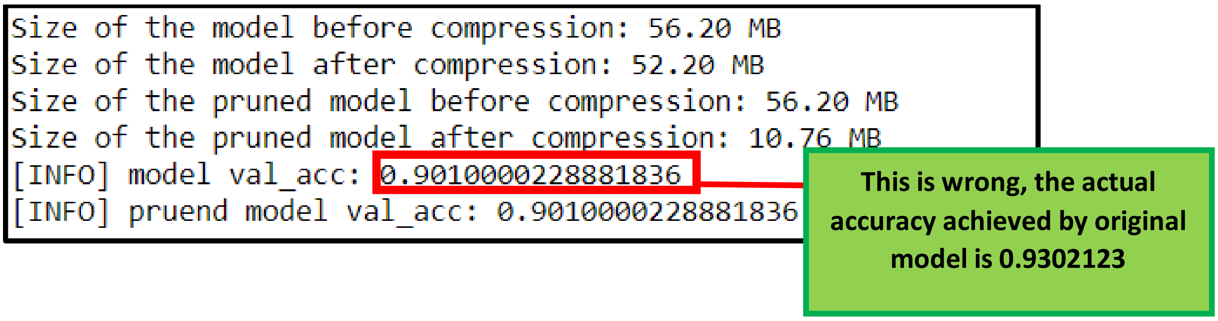
Pruning and model recovering

Training



```
Epoch 00006: ReduceLROnPlateau reducing learning rate to 0.004999999888241291.
250/250 [==============================] - 60s 239ms/step - loss: 0.0271 - acc: 0.9900 - val_loss: 0.2046 - val_acc: 0.9250
Epoch 7/15
250/250 [==============================] - 59s 238ms/step - loss: 0.0660 - acc: 0.9830 - val_loss: 0.1757 - val_acc: 0.9310
Epoch 8/15
250/250 [==============================] - 60s 239ms/step - loss: 0.0517 - acc: 0.9850 - val_loss: 0.2142 - val_acc: 0.9250
Epoch 9/15
249/250 [============================>.] - ETA: 0s - loss: 0.2851 - acc: 0.8830
Epoch 00009: ReduceLROnPlateau reducing learning rate to 0.0024999999441206455.
250/250 [==============================] - 60s 239ms/step - loss: 0.2853 - acc: 0.8830 - val_loss: 0.3198 - val_acc: 0.8440
Epoch 10/15
250/250 [==============================] - 60s 239ms/step - loss: 0.1623 - acc: 0.9460 - val_loss: 0.2478 - val_acc: 0.8950
Epoch 11/15
250/250 [==============================] - 60s 238ms/step - loss: 0.2252 - acc: 0.9115 - val_loss: 0.2465 - val_acc: 0.9030
Epoch 12/15
249/250 [============================>.] - ETA: 0s - loss: 0.2181 - acc: 0.9217
Epoch 00012: ReduceLROnPlateau reducing learning rate to 0.0012499999720603228.
250/250 [==============================] - 60s 239ms/step - loss: 0.2176 - acc: 0.9220 - val_loss: 0.2478 - val_acc: 0.8960
Epoch 13/15
250/250 [==============================] - 59s 238ms/step - loss: 0.2076 - acc: 0.9290 - val_loss: 0.3046 - val_acc: 0.8660
Epoch 14/15
250/250 [==============================] - 60s 239ms/step - loss: 0.1644 - acc: 0.9410 - val_loss: 0.2522 - val_acc: 0.8960
Epoch 15/15
249/250 [============================>.] - ETA: 0s - loss: 0.1408 - acc: 0.9553
Epoch 00015: ReduceLROnPlateau reducing learning rate to 0.0006249999860301614.
250/250 [==============================] - 60s 238ms/step - loss: 0.1407 - acc: 0.9555 - val_loss: 0.2249 - val_acc: 0.9010
```

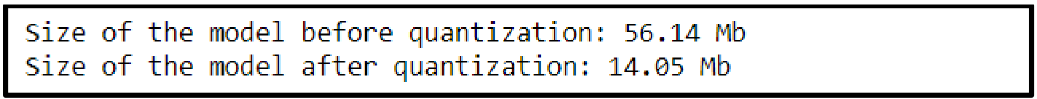
Comparision of compressed model size and validation accuracy



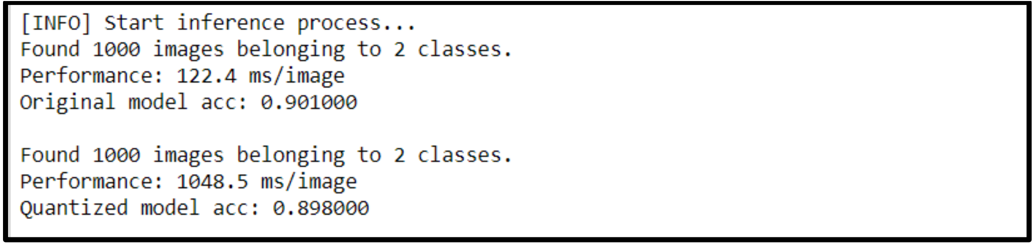
```
Size of the model before compression: 56.20 MB
Size of the model after compression: 52.20 MB
Size of the pruned model before compression: 56.20 MB
Size of the pruned model after compression: 10.76 MB
[INFO] model val_acc: 0.9010000228881836
[INFO] pruend model val_acc: 0.9010000228881836
```

This is wrong, the actual accuracy achieved by original model is 0.9302123

- Quantization



```
Size of the model before quantization: 56.14 Mb
Size of the model after quantization: 14.05 Mb
```



```
[INFO] Start inference process...
Found 1000 images belonging to 2 classes.
Performance: 122.4 ms/image
Original model acc: 0.901000

Found 1000 images belonging to 2 classes.
Performance: 1048.5 ms/image
Quantized model acc: 0.898000
```