

Pruning → Quantization V2



Status

In progress

ResNet50

✓
17m

[75]



```
Original Dense Model Size Model=97.49 MiB
Original Model Validation Accuracy: 90.57114228456913 %
Granular-Magnitude Pruning
Sensitivity Scan Time(mins): 6.4385687987009685
Sparsity for each Layer: {'conv1.weight': 0.30000000000000004, 'layer1.0.
Pruning Time Consumed (mins): 0.00012293656667073568
Total Pruning Time Consumed (mins): 6.438692259788513

Pruned Model has size=8.03 MiB(non-zeros) = 8.24% of Original model size

Pruned Model has Accuracy=10.34 % = -80.23% of Original model Accuracy
```

Comparison Table			
	Original Model	Pruned Model	Quantized Model
Latency (ms/sample)	18.1	18.6	6.8
Accuracy (%)	90.571	79.128	78.146
Params (M)	25.56	2.11	*
Size (MiB)	102.523	102.523	26.145
MAC (M)	86	86	*

Profiling Original and Optimized Model			
Metrics	Original Model	Optimized Model	Speed-Up/Compression Achieved
Latency	18.1	6.8	2.66 x
Parameters	25.56	2.11	91.74 %
Memory	102.523	26.145	74.50 %
MACs(≈ 2FLOPs)	86	86	0.00 %
Accuracy	90.571	78.146	-13.72 %

DenseNet121

===== Comparison Table =====			
	Original Model	Pruned Model	Quantized Model
Latency (ms/sample)	21.0	21.2	13.4
Accuracy (%)	89.709	81.303	78.106
Params (M)	7.98	0.62	0.07
Size (MiB)	32.468	32.468	8.965
MAC (M)	59	59	*

===== Profiling Original and Optimized Model =====			
Metrics	Original Model	Optimized Model	Speed-Up/Compression Achieved
Latency	21.0	13.4	1.57 x
Parameters	7.98	0.62	92.23 %
Memory	32.468	8.965	72.39 %
MACs(≈ 2FLOPs)	59	59	0.00 %
Accuracy	89.709	78.106	-12.93 %

OLD DENSENET (LATENCY ISSUE)

===== Comparison Table =====			
	Original Model	Pruned Model	Quantized Model
Latency (ms/sample)	19.1	24.6	20.8
Accuracy (%)	89.709	81.373	77.766
Params (M)	7.98	0.62	0.07
Size (MiB)	32.468	32.468	8.965
MAC (M)	59	59	*

===== Profiling Original and Optimized Model =====			
Metrics	Original Model	Optimized Model	Speed-Up/Compression Achieved
Latency	19.1	20.8	0.92 x
Parameters	7.98	0.62	92.23 %
Memory	32.468	8.965	72.39 %
MACs(≈ 2FLOPs)	59	59	0.00 %
Accuracy	89.709	77.766	-13.31 %

VGG16

Comparison Table			
	Original Model	Pruned Model	Quantized Model
Latency (ms/sample)	62.4	62.7	18.3
Accuracy (%)	90.571	83.417	27.255
Params (M)	138.36	7.86	*
Size (MiB)	553.439	553.439	138.637
MAC (M)	437	437	*

Profiling Original and Optimized Model			
Metrics	Original Model	Optimized Model	Speed-Up/Compression Achieved
Latency	62.4	18.3	3.41 x
Parameters	138.36	7.86	94.32 %
Memory	553.439	138.637	74.95 %
MACs(≈ 2FLOPs)	437	437	0.00 %
Accuracy	90.571	27.255	-69.91 %

EfficientNetV2

Comparison Table			
	Original Model	Pruned Model	Quantized Model
Latency (ms/sample)	28.8	32.8	25.4
Accuracy (%)	90.321	73.968	57.064
Params (M)	21.46	1.6	*
Size (MiB)	86.677	86.677	23.93
MAC (M)	63	63	*

Profiling Original and Optimized Model			
Metrics	Original Model	Optimized Model	Speed-Up/Compression Achieved
Latency	28.8	25.4	1.13 x
Parameters	21.46	1.6	92.54 %
Memory	86.677	23.93	72.39 %
MACs(≈ 2FLOPs)	63	63	0.00 %
Accuracy	90.321	57.064	-36.82 %

GoogLeNet

```

Original Dense Model Size Model=25.27 MiB
Original Model Validation Accuracy: 40.67134268537074 %
Granular-Magnitude Pruning
Sensitivity Scan Time(mins): 7.50667507648468
Sparsity for each Layer: {'conv1.conv.weight': 0.4000000000000013, 'conv2.conv.weight': 0.5500000000000002, 'conv3.conv.weight': 0.5500000000000002}
Pruning Time Consumed (mins): 0.0001336971918741862
Total Pruning Time Consumed (mins): 7.506819343566894

Pruned Model has size=3.70 MiB(non-zeros) = 14.63% of Original model size

Pruned Model has Accuracy=10.61 % = -30.06% of Original model Accuracy

```

Comparison Table			
	Original Model	Pruned Model	Quantized Model
Latency (ms/sample)	8.3	8.6	4.1
Accuracy (%)	40.671	80.331	80.2
Params (M)	6.62	0.97	*
Size (MiB)	26.654	26.654	6.869
MAC (M)	32	32	*

Profiling Original and Optimized Model			
Metrics	Original Model	Optimized Model	Speed-Up/Compression Achieved
Latency	8.3	4.1	2.02 x
Parameters	6.62	0.97	85.35 %
Memory	26.654	6.869	74.23 %
MACs(≈ 2FLOPs)	32	32	0.00 %
Accuracy	40.671	80.2	97.19 %

Mobilnet v3 large

===== Comparison Table =====			
	Original Model	Pruned Model	Quantized Model
Latency (ms/sample)	14.2	14.5	5.5
Accuracy (%)	82.285	59.359	37.515
Params (M)	5.48	0.86	*
Size (MiB)	22.12	22.12	5.969
MAC (M)	8	8	*

===== Profiling Original and Optimized Model =====			
Metrics	Original Model	Optimized Model	Speed-Up/Compression Achieved
Latency	14.2	5.5	2.58 x
Parameters	5.48	0.86	84.31 %
Memory	22.12	5.969	73.02 %
MACs(≈ 2FLOPs)	8	8	0.00 %
Accuracy	82.285	37.515	-54.41 %