ECE361E Final Project

M2: Model Quantization

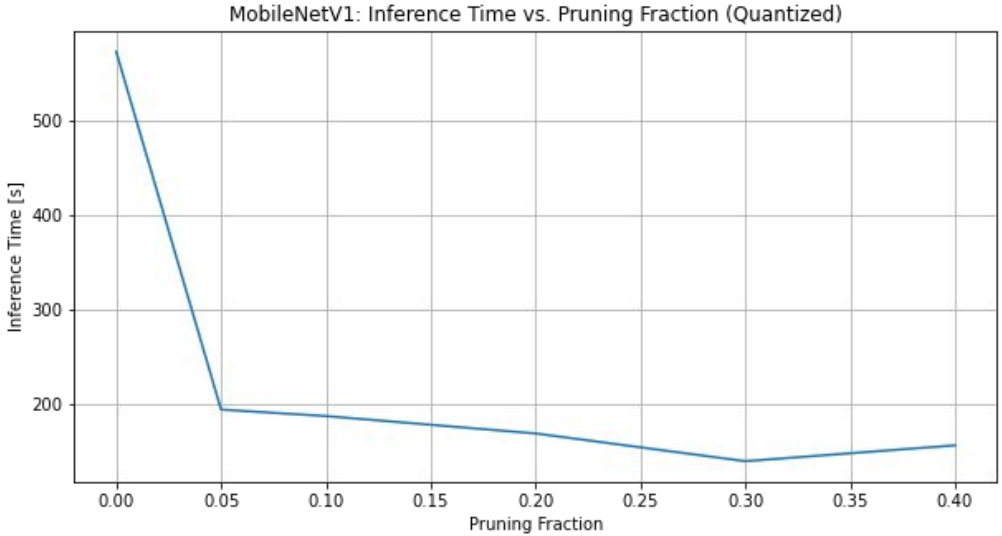
Project Members: Uttami Godha, Jesus Lopez Neira

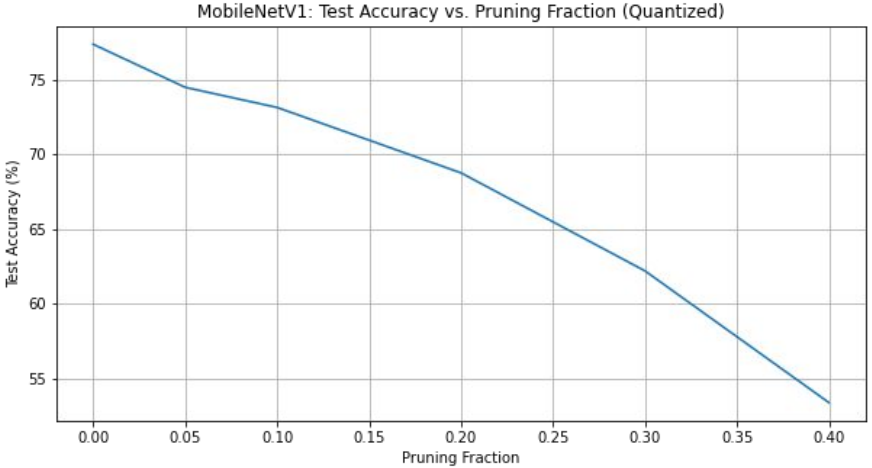
Jesus's tasks- quantization function, exploration
Uttami's tasks- deployment on RPi with ONNX, Table 2 calculations, graphs

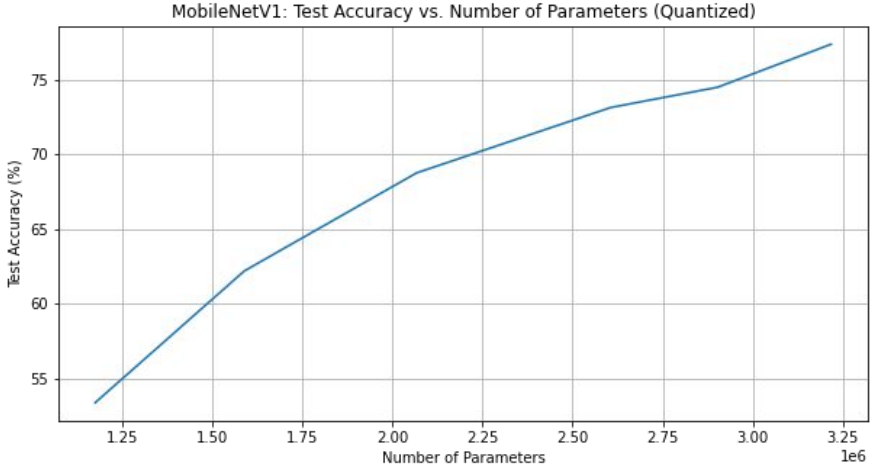
General Approach

In this milestone we compared the efficiency of the top 5 models from M1 using different pruning ratios with quantization, following the methodology:

1. Convert MobileNet to an 8-bit int model using static post-training quantization
2. Deploy models in Raspberry Pi using ONNX
3. Compare models through a results and metrics discussion


MobileNetV1: Inference Time vs. Pruning Fraction (Quantized)
Inference Time [s]
500
400
300
200
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Pruning Fraction


MobileNetV1: Test Accuracy vs. Pruning Fraction (Quantized)
Test Accuracy (%)
75
70
65
60
55
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Pruning Fraction


MobileNetV1: Test Accuracy vs. Number of Parameters (Quantized)
Test Accuracy (%)
75
70
65
60
55
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
3.25
1e6
Number of Parameters

M2.2. Table 2

Pruning fraction	#Fine-tuning epochs	#Params	Max memory usage [MB]	Average latency per image [ms]	Max power consumption [W]	Average energy consumption per image [mJ]
-	-	3217226	81	57.331	4.496	1267.842
0.05	5	2902069	77	19.387	4.481	425.536
0.1	5	2606173	78	18.699	4.481	421.757
0.2	5	2066884	77	16.854	4.470	416.903
0.3	5	1588917	77	13.916	4.439	314.577
0.4	5	1174647	80	15.595	4.366	407.655

Exploration and Conclusions

✓ The addition of quantization to pruned models had a significant impact, particularly on the average latency and average energy consumption per image. This method allowed us to speed up the inference process and save energy to an extent that was not possible with just pruning.

✖ The accuracy suffered a bit after the pruning fraction was increased, which becomes a point of investigation in the final milestone. The memory usage in our case also didn't drop as much as expected.

💡 For further exploration, we plan to do a visualization approach where we can compare the importance of each given layer relative to others.