Milestone 2
Model Quantization

Drew Hardie
Quantized models and collected data

Prathik Srinivasan
Ran inference and collected inference data

Approach

- Static Quantization on existing onnx models
- Converted and quantized a non-pruned mobilenetv1 model as a baseline
- Ran and measured inference results of quantized models

Pruning Fraction	Fine-tuning epochs	parameters	Maximum memory usage [MB] rpi_base=85	average latency per image [ms]	Maximum power consumption [W]	Average energy consumption per image [mJ]
0.05	5	2,891,687	45	25.71	6.68	337.41
0.1	5	2,596,342	42	28.99	6.69	369.11
0.2	5	2,058,145	40	21.35	6.68	277.37
0.3	5	1,581,271	39	18.01	6.72	232.77
0.4	5	1,168,093	35	19.90	6.68	244.99
0.5	5	818,252	33	10.45	6.72	137.69

*correction to M1 calculations

Results

Pruning Fraction	Fine-tuning epochs	parameters	Maximum memory usage [MB] rpi_base=85	average latency per image [ms]	Maximum power consumption [W]	Average energy consumption per image [mJ]	M2 Test Accuracy
0	5	3,206,464	38	9.59	7.36	127.00	77.52%
0.05	5	2,891,867	44	8.17	7.33	114.92	74.35%
0.1	5	2,596,522	37	7.69	7.29	107.24	73.02%
0.2	5	2,058,325	35	6.10	7.20	86.14	69.64%
0.3	5	1,581,451	36	5.63	7.15	77.20	63.18%
0.4	5	1,168,273	33	4.57	7.05	62.88	51.21%

Results

After Quantization:

- +180 parameters
- Slight decrease of max memory usage
- ≈+0.5W on average for max power consumption
- Average Latency ≈3.6x faster
- Average Energy Consumption ≈3.3x lower

M1 Test Accuracy	M2 Test Accuracy	Difference	Pruning Fraction
77.68%	77.52%	-0.16	0.0
74.63%	74.35%	-0.28	0.05
73.28%	73.02%	-0.26	0.1
69.68%	69.64%	-0.04	0.2
63.15%	63.18%	+0.03	0.3
51.63%	51.21%	-0.42	0.4

Conclusion

- Quantization brings significant improvements to latency and energy consumption
- Different architecture?
- Image preprocessing?