ECE361E Final Project

Sidharth Babu, Tianda Huang

Approach

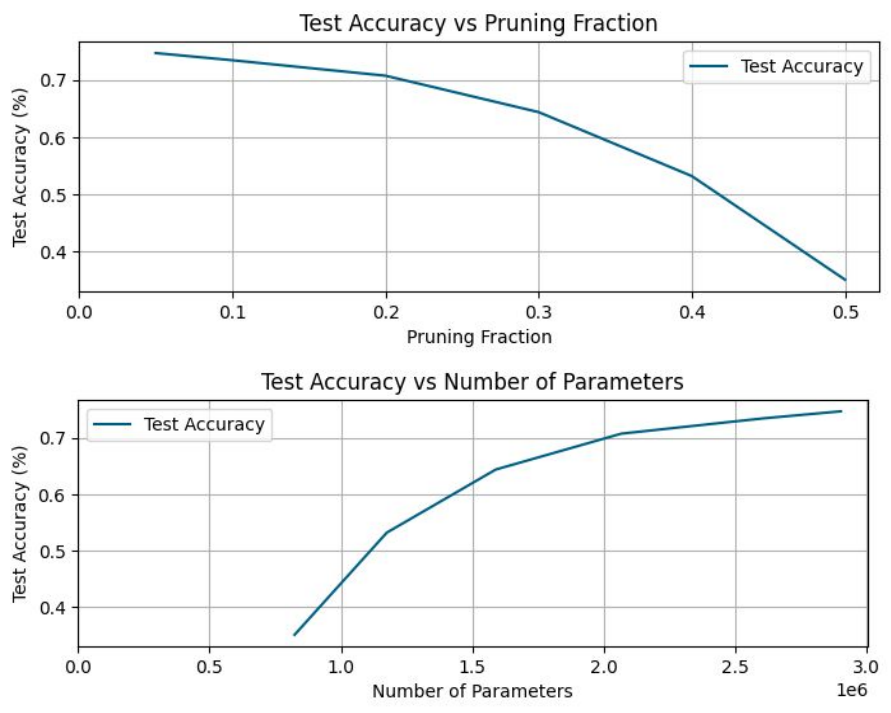
Required Portion

- Trained MobileNetv1 using Torch and deployed using ONNX
- Repurposed and tuned Demo pruning code to accelerate development of our pruning code
- During deployment, took holistic measurements to help build intuition for later project items.

Exploration Portion

- We read multiple papers on quantization
- We've researched frameworks and tools to use
- We've researched model architecture alternatives

Accuracy after Pruning



Statistics

Pruning Fraction	Fine-tuning epochs	Number of Parameters	Max RAM usage (Mb)	Inference time per image (ms)	Max power consumption (W)	Average energy per image (J)
0.05	5	2902069	138.0	40.42522995471954	6.706	0.24955627650519782
0.1	5	2606173	135.0	37.9954131603241	6.512	0.23470382227730077
0.2	5	2066884	132.0	29.80535206794739	6.542	0.1855600232613007
0.3	5	1588917	130.0	26.49245958328247	6.588	0.16303766687611596
0.4	5	1174647	128.0	18.59753966331482	6.576	0.11552728127113261
0.5	5	823722	127.0	12.149637842178345	6.576	0.07667170938833152

M1.5

Conclusions

- Pruning is very effective
- Needs to be combined with other steps in order to enhance performance
- Other architectures will likely need to be explored
- Experimenting with different environments and frameworks
- Researching the target hardware