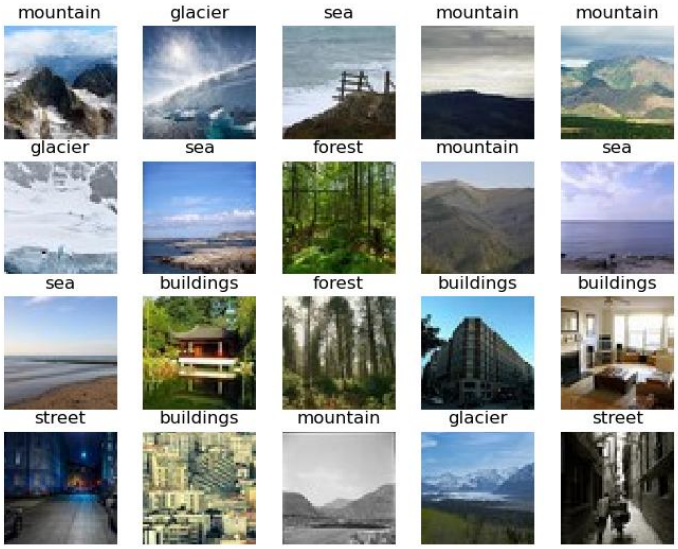
Micro-controller Image Classification

Tools and skills: C/C++, Python, TensorFlow, STM32 Micro-controller, Data analysis

Julian Lapenna
7jlapenna@gmail.com
+1 647 850 7691



Links: Course Information

Goal:

Deployed a machine learning model on an STM32 Microcontroller to optimize inference speed, performance accuracy, and power consumption.

Outcome:

Machine learning model predicts landscape images from six classes with 80% accuracy and with a time for inference of 41.1ms (~24.32 inferences/second). Model runs on STM32 Cortex M4 Core.

Key Features:

Machine learning model:

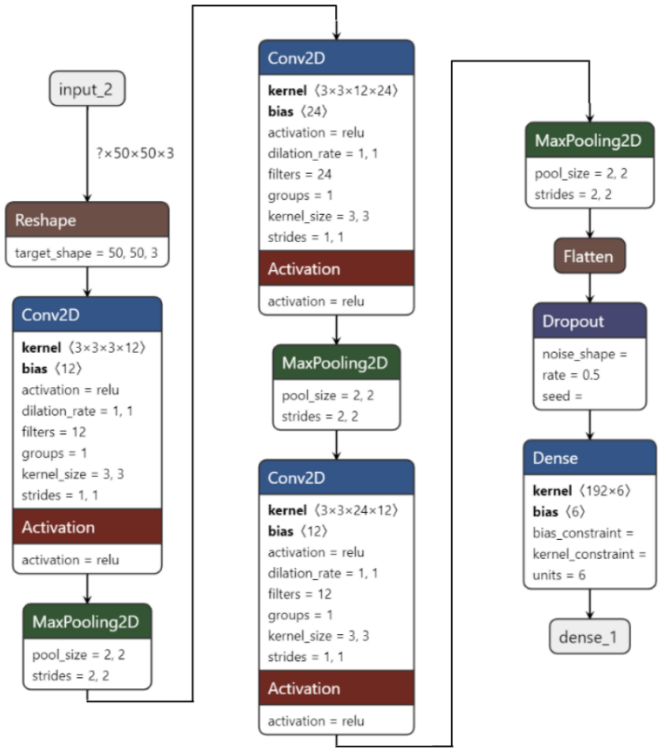
- Built and trained a convolutional neural network (CNN) using TensorFlow, outputting landscape classification
- Tested model architectures for optimal performance of 80% accuracy with minimal Flash and RAM usage

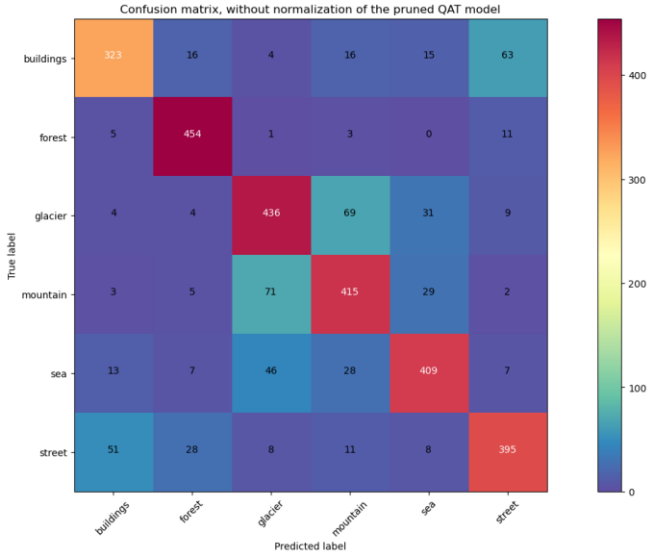
Hardware analysis:

- Quantized and pruned the model to reduce the size and memory usage for efficient inference
- Model size of 563 kB (71% of board Flash) and RAM requirement of 68 kB (55% of board SRAM)

Optimization:

- Used a Prune Quantize-Aware Trained (PQAT) model for Flash reduction of 3x while maintaining <1% performance loss





Confusion matrix of PQAT model (above). Accuracy vs size comparison of different model optimizations, size expressed as radius (below).