HW03: weight quantization for CNNs

AI graduate school

20204065 Heejun Jung

1. Accuracy (CIFAR10)

FP32: 92.04%, DA: QAT: 91.51%

1. The advantage of weight-only quantization is to reduce inference time and reduce memory saving space. Besides, static quantization, weight-only quantization, is well preserve model performance than dynamic quantization that quantize weight and activation after model training.
2. Hyper-parameters for QAT

Optimizer: SGD, number of bits: 16, weight decay: 2e-5, learning rate: 1e-1, epoch: 120, seed: 100

1. QAT
   1. Grid search was conducted from {4, 8, 16} to find out best case of step size
   2. Grid search was implemented for learning rate from {1e-1, 1e-2, 1e-3} to find best case then set again from {2e-1, 3e-1, 4e-1, 5e-1} and set epoch from {60, 10}, batch size from {256, 512}. In the last, weight decay was set from {1e-5 ,2e-5, …, 5e-5}.