Training and Pruning of Convolutional Neural Network in Winograd domain

Swastik Sahu, Master of Science in Computer Science

University of Dublin, Trinity College, 2019

Supervisor: Dr. David Gregg

Deep convolutional neural networks can take a very long time to train on a large data-set. In critical systems like autonomous vehicles, low latency inference is desired. Most IoT and mobile devices are not powerful enough to train a successful predictive model using the data that's accessible to the device. The success of convolutional neural networks in these situations is limited by the compute power. Using Winograd's minimal filtering algorithms, we can reduce the number of multiplication operations needed for the convolution operation in spatial domain. Convolution in Winograd domain is faster when performed with small input patch and small filter kernels.

Use of Winograd techniques to perform the convolution is popular in deep learning libraries but the weights are transformed back to spatial domain after the convolution operation. In this research, we have written CPU and GPU implementations to train the weights in the Winograd domain and used different CNN architectures to train our model in spatial domain and in Winograd domain separately, on MNIST and CIFAR-10 data-set. Our GPU implementation of Winograd convolution is $\sim 2\times$ times faster than convolution in spatial domain for MNIST data-set, and $\sim 3.4\times$ times faster for CIFAR-10 data-set. Higher accuracy levels and quicker convergence was observed while training in Winograd domain for the MNIST data-set. For CIFAR-10 data-set, the effective time to converge when training in Winograd domain was $\sim 2.25\times$ times faster compared to training in spatial domain. After training separately in spatial domain and in Winograd domain, the respective weights were pruned in an iterative manner. The accuracy drop for weights trained in Winograd domain was observed to be lower than that observed for weights trained in spatial domain.