Efficient cuDNN-compatible Convolution-Pooling on the GPU

Shunsuke Suita[1], Takahiro Nishimura[1], Hiroki Tokura[1], Koji Nakano[1], Yasuaki Ito[1], Akihiko Kasagi[2], Tsuguchika Tabaru[2]
[1]Department of Information Engineering, Hiroshima University
Higashi-Hiroshima, Japan
[2]Fujitsu Laboratories Ltd.
4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki, Kanagawa, Japan
nakano@cs.hiroshima-u.ac.jp

The main contribution of this paper is to show efficient implementations of the convolution-pooling in the GPU, in which the pooling follows the multiple convolution. Since the multiple convolution and the pooling operations are performed alternately in earlier stages of many Convolutional Neural Networks (CNNs), it is very important to accelerate the convolution-pooling. Our new GPU implementation uses two techniques, (1) convolution interchange with direct sum, and (2) conversion to matrix multiplication. By these techniques, the computational and memory access cost are reduced. Further the convolution interchange is converted to matrix multiplication, which can be computed by cuBLAS very efficiently. Experimental results using Telsa V100 GPU show that our new GPU implementation compatible with cuDNN for the convolution-pooling is 1.34-9.49 times faster than the multiple convolution and then the pooling by cuDNN, the most popular library of primitives to implement the CNNs in the GPU.

Keywords: Deep Learning, Neural Networks, Convolution, Average Pooling, GPU.