Comparative study of CUDA GPU implementations in Python with the fast iterative shrinkage-thresholding algorithm for LASSO

Younsang Cho
Department of Statistics, Inha University

Abstract

A general-purpose GPU (GPGPU) is employed in a variety of domains, including accelerating the spread of deep natural network models; however, further research into its effective implementation is needed. When using the compute unified device architecture (CUDA), which has recently gained popularity, the situation is analogous to use the GPUs and its memory space. This is due to the lack of a gold standard for selecting the most efficient approach for CUDA GPU parallel computation. Contrarily, as solving the least absolute shrinkage and selection operator (LASSO) regression fully consists of the basic linear algebra operations, the computation using GPGPU is more effective than other models. Additionally, its optimization problem often requires fast and efficient calculations. The purpose of this study is to provide brief introductions to the implementation approaches and numerically compare the computational efficiency of GPU parallel computation with that of the fast iterative shrinkage-thresholding algorithm for LASSO. This study contributes to providing gold standards for the CUDA GPU parallel computation, considering both computational efficiency and ease of implementation. Based on our comparison results, we recommend implementing the CUDA GPU parallel computation using Python, with either a dynamic-link library or PyTorch for the iterative algorithms.

Keywords: Compute unified device architecture, Graphics processing unit, Fast iterative shrinkage-thresholding algorithm, LASSO, Python