Two-level Group Convolution

Youngkyu Lee\textsuperscript{1}, Jongho Park\textsuperscript{2} and Chang-Ock Lee\textsuperscript{1}

1) \textit{Department of Mathematical Sciences, KAIST, Daejeon 34141, KOREA}
2) \textit{Natural Science Research Institute, KAIST, Daejeon 34141, KOREA}

Corresponding Author: Chang-Ock Lee, colee@kaist.edu

ABSTRACT

Group convolution \cite{1} has been widely used in order to reduce the computation time of convolution, which takes most of the training time of convolutional neural networks. However, it is well known that a large number of groups significantly reduce the performance of group convolution. In this presentation, we propose a new convolution methodology called “two-level” group convolution that is robust with respect to the increase of the number of groups and suitable for multi-GPU parallel computation. We first observe that the group convolution can be interpreted as a one-level block Jacobi approximation \cite{2} of the standard convolution, which is a popular notion in the field of numerical analysis. In numerical analysis, there have been numerous studies on the two-level method that introduces an intergroup structure that resolves the performance degradation problem without disturbing parallel computation. Motivated by these, we introduce a coarse-level structure which promotes intergroup communication without being a bottleneck in the group convolution. We show that all the additional work induced by the coarse-level structure can be efficiently processed in a distributed memory system. Numerical results that verify the robustness of the proposed method with respect to the number of groups are presented. Moreover, we compare the proposed method to various approaches for group convolution in order to highlight the superiority of the proposed method in terms of execution time, memory efficiency, and performance.

REFERENCES