論文 / 著書情報
Article / Book Information

題目(和文)	大規模畳み込みニューラルネットワークの階層的なハイブリッド並列学習
Title(English)	Hierarchical Hybrid Parallel Training of Large-Scale Convolutional Neural Networks
著者(和文)	大山洋介
Author(English)	Yosuke Oyama
出典(和文)	学位:博士(理学), 学位授与機関:東京工業大学, 報告番号:甲第11892号, 授与年月日:2021年3月26日, 学位の種別:課程博士, 審査員:松岡 聡,増原 英彦,遠藤 敏夫,脇田 建,横田 理央
Citation(English)	Degree:Doctor (Science), Conferring organization: Tokyo Institute of Technology, Report number:甲第11892号, Conferred date:2021/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	数理・計算科学 数理・計算科学	系 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of	（ 理学 ）
学生氏名： Student's Name	大山洋介		指導教員（主）： Academic Supervisor(main)	松岡聡	
			指導教員（副）： Academic Supervisor(sub)		

要旨（英文 800 語程度）
Thesis Summary（approx.800 English Words ）

In the last decades, deep learning technology has attracted substantial research interests. Deep learning has been empowered by the advance of network architecture and training algorithms, such as the growth of available data to train deep neural networks and the increase of the computation capability of high-performance GPUs and supercomputers. Specifically, many studies successfully have adopted data-parallelism to distribute their training workloads among dozens and even hundreds of accelerators due to its simple parallelization design. As the demand for training deep learning models with complex and massive data increases continuously, these improvements must take place in parallel to keep up research speed. However, trends in hardware, software, and model architecture for deep learning are changing rapidly. For example, recent accelerators equip specialized hardware in matrix multiplication operations frequently used in deep learning, but their performance is not fully investigated for various deep learning workloads. Another change is the emergence of high-resolution, large-scale models that perform end-to-end learning on scientific data, which cannot be trained with conventional data-parallel methods. For these reasons, there is a strong demand for a general-purpose way of accelerating and saving memory for more diverse model architectures, computational precisions, and large-scale models on multiple levels of parallelism.

In this thesis, we propose acceleration algorithms that maximize the parallel efficiency of CNNs at two different levels, intra-processor, and inter-processor parallelism. For intra-processor optimization, we present the μ-cuDNN library, which applies loop optimization and adaptively uses various algorithms and computational precisions for convolution kernels. This library replaces the cuDNN library transparently to optimize the convolution performance, the de-facto standard kernel library for deep learning frameworks. Thus, our library is widely applicable to such frameworks. Since convolution is one of the most computationally-intensive parts of deep learning training and inference, accelerating convolutional computation plays a crucial role in accelerating such jobs in many fields. Our loop optimization method allows users to select a broader range of convolutional algorithms without changing computation semantics. We combine the μ-cuDNN library into two frameworks, Caffe and TensorFlow, and show that it achieves reasonable performance improvements in several different CNNs; we achieve speedups of 1.60x for AlexNet and 1.30x for ResNet-18 on an NVIDIA V100 GPU. We also show that μ-cuDNN achieves speedups of up to 4.54x, and 1.60x on average for DeepBench convolutional layers, and demonstrate that NVIDIA GPU's single-precision arithmetic units are still beneficial to accelerate convolution on half-precision float data. These results indicate that micro-batching can seamlessly increase deep learning performance while using the same overall memory footprint. Moreover, we propose an interface to use this information for multi-node training utilizing the property that μ-cuDNN can obtain layer parameters via the cuDNN interface. Furthermore, we show that we can combine the ONNX data format with our algorithm to apply the loop optimization algorithm without changing its framework. We also discuss whether μ-cuDNN's algorithm can be extended to layer types other than convolution and different memory layouts.

For inter-processor optimization, we present an end-to-end hybrid-parallel training approach for strong-scaling training large-sample 3D CNNs, which applies spatial partitioning. Since data-parallel training frameworks cannot train large models beyond the memory capability of a single GPU, this approach enables users to improve the inference accuracy of such CNNs by increasing the input dimensions. Specifically, we propose various techniques to optimize the performance for better scalability; we show GPU kernels designed for 2D CNNs can be inefficient. Thus, the implementation of custom kernels for high-dimensional layers is necessary. Moreover, we propose a spatial-partitioned sample I/O method to mitigate the data read overhead, which is essential to achieve practical strong-scaling, such as spatial-partitioned sample I/O. We demonstrate training on full-resolution data samples for the CosmoFlow

network (512^3) and the 3D U-Net (256^3). Our performance results show good strong- and weak- scaling on up to 2048 NVIDIA V100 GPUs; we achieve 1.77x of speedup on 2048 GPUs over 512 GPUs with the same mini-batch size of 64 for the CosmoFlow network and achieve 1.42x of speedup on 512 GPUs over 256 GPUs with the same mini-batch size of 16 for the 3D U-Net. We also propose a performance model for the hybrid-parallel training framework to demonstrate that its performance is predictable with benchmark results using a limited number of compute nodes. Besides, we present a significant improvement in prediction accuracy by using full-resolution data of the CosmoFlow cosmological data.
This thesis provides a means to accelerate the training of CNNs across multiple levels of parallelism significantly through our proposed approaches. We expect this to discover new scientific knowledge by training large-scale, high-dimensional, and high-resolution CNNs in highly parallel environments.