

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)	審査の要旨
Type(English)	Exam Summary

論文審査の要旨及び審査員

報告番号	甲第 号			学位申請者氏名		大山 洋介	
論文審査 審査員		氏 名	職 名		氏 名	職 名	
	主査	松岡 聡	特任教授	審査員	横田 理央	准教授	
	審査員	増原 英彦	教授				
		遠藤 敏夫	教授				
		脇田 建	准教授				

論文審査の要旨（2000 字程度）

本論文は「Hierarchical Hybrid Parallel Training of Large-Scale Convolutional Neural Networks（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）」と題し、深層学習で使用されるモデルの一種である畳み込みニューラルネットワーク（CNN）に対し、単一アクセラレータ内と複数アクセラレータ間という異なる階層で利用可能な並列性に対してアルゴリズムを最適化することで学習を高速化する手法を提案するもので、英文で全 7 章から構成されている。

第 1 章「Introduction」では、本論文の導入として、近年の深層学習の研究におけるハードウェア・ソフトウェア・問題サイズに関する問題を述べ、特に計算量が大きな問題となる CNN について、これらの性能改善が深層学習研究において大きな意義があることを説明している。また、このようなモデルの並列性を用いて性能最適化を行う提案手法の概要を説明し、最後に本論文の構成を述べている。

第 2 章「Background」では、本論文の背景として、深層学習の基礎的事項について説明し、そのような深層学習モデルの学習・推論を高性能計算技術を用いて行う手法について説明している。

第 3 章「Automatic optimization of computational kernels」では、CNN で計算時間上のボトルネックとなる畳み込みレイヤーの計算を最適化する手法を提案している。具体的には、単一 GPU 内でワークスペースメモリ制約を維持したまま、動的計画法を用いた最外ループの分割によって使用する畳み込みのアルゴリズムを最適化する手法を提案している。また、このアルゴリズムがワークスペースメモリの削減に寄与するだけではなく、混合精度演算をとまなうケースでもより高速でメモリ使用量の少ないアルゴリズムを選択できることもあわせて説明している。NVIDIA 社の Volta 世代の GPU を用いた評価では、提案手法が DeepBench マイクロベンチマークについて平均 1.60 倍の高速化を、ResNet-18 の学習については 1.30 倍の高速化を達成することを確認している。

第 4 章「Training 3D CNNs with hybrid-parallelization」では、メモリ使用量が非常に大きく通常のデータ並列学習手法ではモデルサイズがスケールしない 3 次元 CNN について、空間方向にモデルを分割してハイブリッド並列で学習を行う手法を提案している。計算上のボトルネックが未知である巨大なモデルについて計算カーネル・I/O・推論精度の異なる観点から評価を行い、スケーラビリティを維持するためにそれぞれについて性能最適化を行っている。また、計算時間に関する性能モデルを提案し、データ並列学習と比較して計算性能が予測しづらいハイブリッド並列学習の性能予測についても性能モデリング手法が有効であることを示している。Lassen スーパーコンピュータを用いた評価では、CosmoFlow モデルおよび 3D U-Net モデルについて最大 2048 GPU を用いて評価を行い、かつ CosmoFlow モデルを用いた評価において、ハイブリッド並列によって初めて可能になる 3 次元 CNN の高解像度化がモデルの推論精度の改善に寄与することを確認している。

第 5 章「Related work」では、第 3 章および第 4 章で提案した 2 つの手法のそれぞれまたは両方に類似した先行研究を述べ、提案手法が各階層において CNN の計算性能を向上させる上で優位性または親和性があることを述べている。

第6章「Discussion」では、提案手法による深層学習の研究やコミュニティに対する貢献について述べ、さらに第3章・第4章で取り扱ったもの以外のモデルやアーキテクチャに提案手法を応用した場合に予想される計算性能について議論している。

第7章「Conclusion」では、本論文の総括を述べている。

以上のように、本研究は深層学習モデルの計算の並列計算機の異なる階層において高速化する手法について提案し、また実験においてその有効性を確認しており、理学上貢献するところ大である。よって、本論文は博士（理学）の学位論文として十分価値があるものと認める。

注意：「論文審査の要旨及び審査員」は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。