様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月　 ５ 日現在

機関番号：１２６０８
研究種目：挑戦的研究（開拓）
研究期間：2020～2023
課題番号：２０Ｋ２０６２４

研究課題名（和文）超巨大ニューラルネットの継続学習への型破りな線形代数技術の適用

研究課題名（英文）Application of Unconventional Linear Algebra Techniques to Continuous Learning in Supergiant Neural Networks

研究代表者

横田　理央（Yokota, Rio）

東京工業大学・学術国際情報センター・教授

研究者番号：２０７６０５７３

交付決定額（研究期間全体）：（直接経費）　　19,500,000円

研究成果の概要（和文）：深層継続学習ではFisher情報行列の行列分解を用いることで性能が向上することが示されている。しかし、Fisher情報行列はパラメータ数Nの2乗の要素数を持つ密行列であるため、そのまま行列分解を直接行うことが困難である。本研究では階層的低ランク近似法であるH^2行列を用いることで、この計算量をO(N)に低減した。さらに、fill-inブロックを予め計算し共有基底に含めてULV分解を行うことでに全ての対角ブロックを並列に処理する手法を提案した。また、テンソルコアのような低精度演算器でも悪条件の行列の分解ができるよう、精度を補正する手法を開発した。

研究成果の学術的意義や社会的意義
Fisher情報行列は継続学習やモデル・マージング、連合学習を行う際に有用であることが知られているが、その計算コストは膨大でありモデルの規模が近年急激に増大していることからも、その計算を高速化する手法が求められている。これまでKronecker因子分解による近似を行うことでO(N^1.5)の計算量にする方法が提案されているが、本研究ではこれをO(N)にまで低減できたことは意義深い。これにより、継続学習、モデル・マージング、連合学習の研究が加速すれば、一部の限られた大企業の専売特許となっている大規模な生成モデルの構築が、より多くの研究者の共同作業によって分担して構築できるようになる。

研究成果の概要（英文）：It has been shown that using matrix factorization of the Fisher information matrix improves the performance of continual deep learning. However, it is difficult to perform matrix factorization directly on the Fisher information matrix because it is a dense matrix where the number of elements grows with the square of the number of parameters N. In this study, we use the H^2 matrix, which is a hierarchical low-rank approximation method that can reduce computational complexity to O(N). Furthermore, we proposed a method to process all diagonal blocks in parallel by performing ULV decomposition with fill-in blocks pre-computed and included in the shared basis. We also developed a method for recovering the numerical accuracy when using low-precision arithmetic units such as tensor cores, which allows us to factorize ill-conditioned matrices.

研究分野：高性能計算

キーワード：階層的低ランク近似法　深層学習　行列分解　テンソルコア

１．研究開始当初の背景

（１）本研究を開始した 2020 年は、個々のタスクに特化した小規模なモデルを皆が学習する時代から、あらゆるタスクに対応できる大規模なモデルを大量の計算資源をもつ一部の者が学習する時代へと変わっていく転換期であった。

（２）研究開始当初は継続学習が鍵となると思われていたが、2024 年時点では GPT-4 や Gemini などのモデルは依然として一から事前学習を行っており、大規模モデルに最新情報を継続的に学習させる仕組みは確立されていない。

２．研究の目的

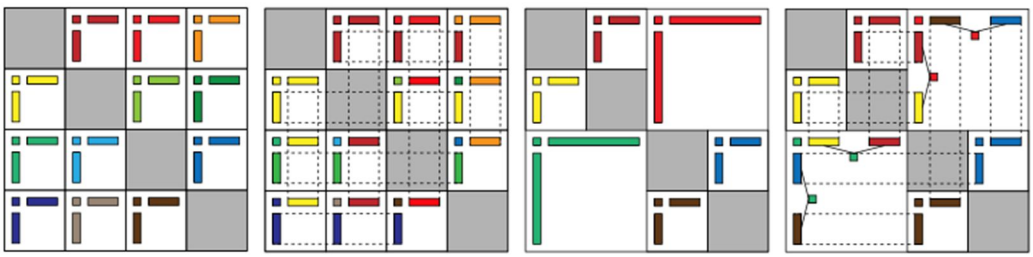
（１）本研究では、超巨大ニューラルネットの継続学習のための型破りな線形代数技術の開発を目的とする。具体的には、分散並列深層学習の収束性の向上、継続学習における致命的忘却の抑制、ハイパーパラメータの予測などを行うことにより、超巨大ニューラルネットの継続学習における課題の解決を図る。

（２）収束性向上させることができる二次最適化で用いられるヘッセ行列やフィッシャー行列、継続学習における致命的忘却を抑制するのに用いられるフィッシャー行列、学習率やバッチサイズなどのハイパーパラメータの予測に用いられるヘッセ行列は、パラメータ数の次元を持つ巨大な密行列であるため、それによる前処理は膨大な計算コストを要する。そこで、本研究では階層的低ランク近似行列(H 行列)を用いてこの計算コストを大幅に低減することを目的とする。

３．研究の方法

（１）これまで深層学習分野ではフィッシャー行列の計算コストがあまりにも膨大であることから対角近似が用いられてきたが、これではあまりにも近似誤差が大きくフィッシャー行列を用いることによる明確な効果は得られていない。研究代表者らはこれまでに、フィッシャー行列のクロネッカー因子分解を用いることでその計算量を $O(N^3)$から $O(N^{3/2})$に低減し、対角近似に比べて大きな効果が得られることを示した。本研究では、H 行列を用いることで任意の近似精度を実現しながらも計算量を O(N)にまで低減できることを示す。

（２）これまでの提案されてきた H 行列の行列分解では並列度の高い $O(N^{3/2})$の手法か並列度の低い O(N)の手法のいずれかであった。本研究が対象とする行列は数百万から数億次元にもなるため分散並列化が必須であり、かつ O(N)でなければ現実的な時間で計算することは困難である。そこで、本研究では高い並列度と O(N)の両方の性質を併せ持つ密行列分解の手法を開発する。



(a) BLR　(b) BLR (uniform basis)　(c) H-matrix/HODLR　(d) H^2-matrix/HSS

図１　４種類の異なる階層的低ランク近似行列

４．研究成果

（１）階層的低ランク近似法には H 行列の他にも基底を共有する H^2 行列や対角ブロックのみを分割する HSS 行列、階層的になっていない BLR などがある。図１に４種類の異なる階層的低ランク近似行列を示す。最も単純な構造は(a)の BLR であり密行列をブロック分割して非対角ブロックを低ランク近似する。(a)ではそれぞれの低ランクブロックの基底が独立なのに対して(b)の構造では基底が同じ行や列のブロックの間で共有されている。(c)のように階層的にブロック分割するもののうち対角ブロックのみを再帰的に分割するものを HODLR 非対角ブロックも再帰的に分割するものを H 行列という。(d)は(c)の構造において(b)のように基底を共有するもので、HODLR の基底を共有するものを HSS、H 行列の基底を共有するものを H^2 行列という。

H 行列では行列分解の際に処理の依存関係のために並列化効率が低下し、HSS 行列では非対角ブロックのランクが増大するため、H^2 行列に比べて高い性能を得ることが難しい。HSS 行列の既存研究では ULV 分解を用いることで処理の依存関係を解消し、全ての対角ブロックを並列に処理する手法が提案されている。しかし、H^2 行列に ULV 分解を適用すると fill-in ブロックの再圧縮の際に共有基底の更新が必要になり、H 行列と同様の依存関係の問題が生じる。本研究では、fill-in ブロックを予め計算し共有基底に含めて ULV 分解を行うことで HSS 行列のように全ての対角ブロックを並列に処理する手法を提案した。

提案手法の概要を図２に示す。まず、密行列をブロック分割し、低ランク近似できない Dense ブロックと低ランク近似できる Low-rank ブロックに分ける。次に、Dense ブロックの行列分解を行い Fill-in するブロックを保存する。これらの Fill-in ブロックと Low-rank ブロックを合わせたものから共有基底を計算する。最後にこの Fill-in も含む共有基底で ULV 分解された緑のブロックに対して行列分解を行う。こうすることで、fill-in ブロックの再圧縮を共有基底を更新することなく行うことができ、H^2 行列の行列分解の課題であった依存関係を解消できる。その結果、H^2行列の本来の長所である O(N)の計算量と HSS-ULV の並列度を併せ持つ H^2-ULV 分解を実現することができた。

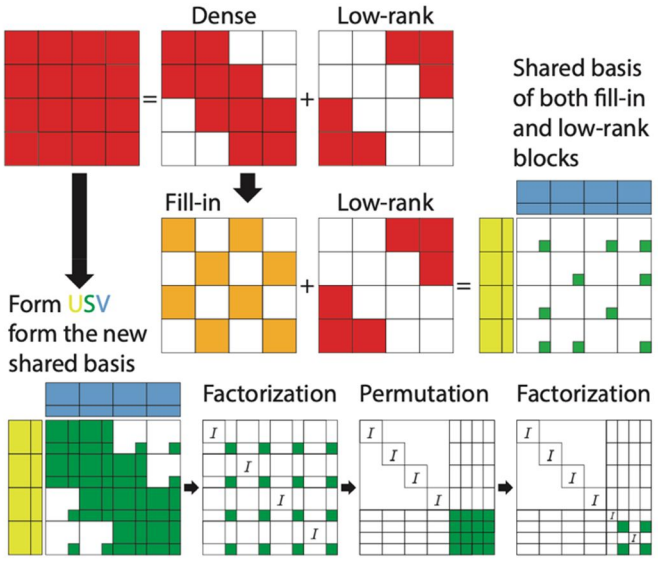


図２　H^2-ULV 分解の概要

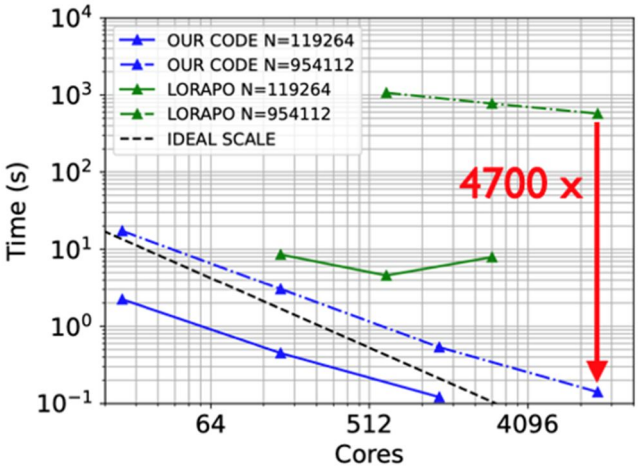


図３　提案手法と LORAPO との比較

提案手法を SC21 の Gordon Bell ファイナリストのコード LORAPO と比較したものを図３に示す。LORAPO は BLR を採用しており O(N^2)の計算量をもつ。一方、提案手法は H^2 行列を採用しているため O(N)の計算量となり今回実験に用いた 100 万次元規模の行列では圧倒的に優位となる。また、Fill-in ブロックを予め共有基底に含めることにより依存関係を解消したことでコア数を増やしたときの並列化効率も圧倒的に良くなっていることが分かる。この成果は高性能計算分野のトップカンファレンスである SC22 に採択された[1]。

（２）上記の結果は CPU 上でのものであるが、本研究ではさらにこれをマルチ GPU 環境で高速に動作するように実装した[2]。本手法の内部では無数の小規模な密行列同士の積が計算され、その GPU 上での実行効率が高速化のカギとなる。このような計算においては Batched GEMM のライブラリをそのまま用いることができるため、疎行列の計算などとは異なり高い FLOP/s を実現できる。ここで、（１）で述べた依存関係の解消が重要な役割を果たす。通常の行列分解では処理に依存関係があるため、Batched GEMM を用いたバッチ処理ができない。しかし、（１）で述べた提案手法を用いることでこの依存関係が解消され、GPU のコアが全てフル稼働するほどの並列度が生まれる。これにより、行列分解を GPU を用いて大幅に高速化することができた。ただし、行列分解が高速化されたことで依存関係が解消できていない前進後退代入の部分の計算時間が目立つようになった。そこで、前進後退代入もブロック間の依存関係なく処理できる手法を開発した。図４に先ほどと同様の LORAPO との比較を示す。提案手法は GPU 上で実行されることで計算時間が大幅に短縮できていることが分かる。このような強スケーリングの実験において 128 台の GPU まで速度が向上し続けていることは本手法の高い並列度を示している。

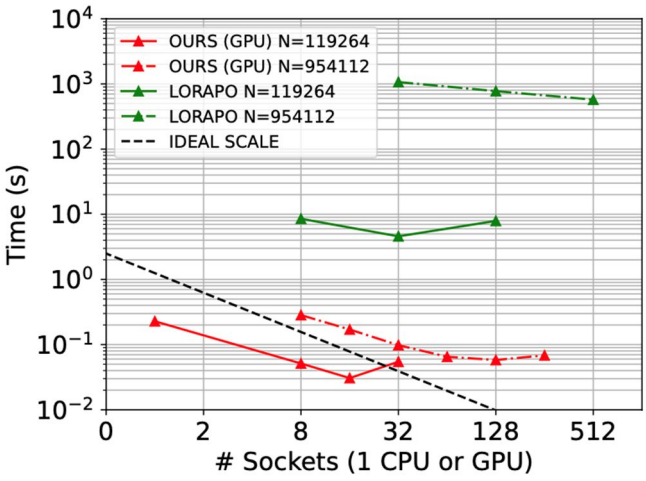


図４　提案手法の GPU 実装と LORAPO との比較

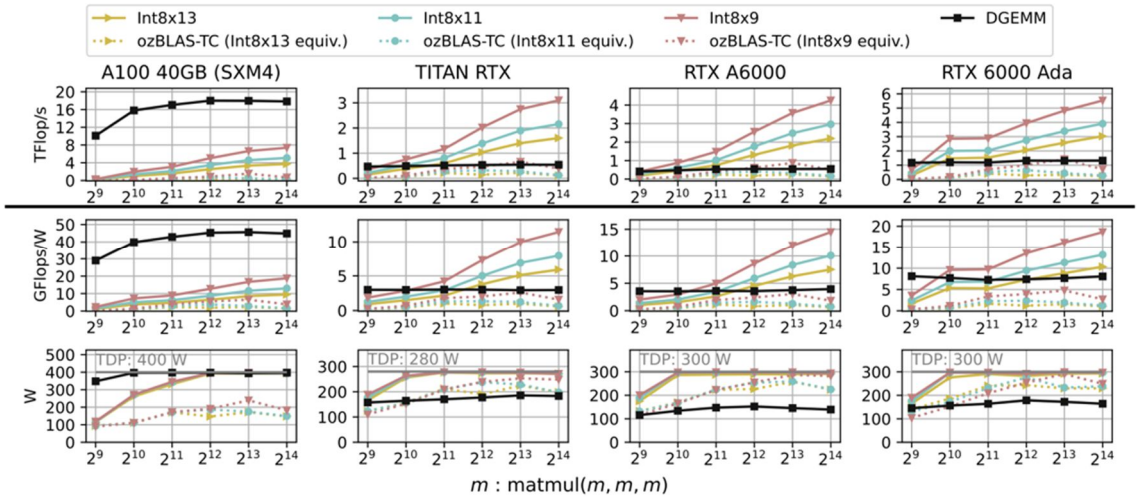


図５　様々な GPU 上での提案手法の演算性能(TFlop/s)と電力効率(GFlops/W)

（３）階層的低ランク近似によるアルゴリズム的な高速化とマルチ GPU 実装によるハードウェア的な高速化により、100 万次元のフィッシャー行列の分解が現実的な時間で行えるようになったが、GPU 上に搭載されているテンソルコアを用いることで更なる高速化が実現できる。ただし、テンソルコアでは入力行列を 16bit や 8bit の低精度型に変換するため、悪条件の行列を分解することができない。そこで、本研究ではテンソルコアのような低精度演算器でも悪条件の行列の分解ができるよう、精度を補正する手法を開発した[3]。図５に 8bit 整数型のテンソルコアを用いながらも倍精度と同等の精度まで回復したときの演算性能および電力効率を示す。倍精度のテンソルコアが搭載されている A100 では提案手法は DGEMM と比べて低い演算性能となっているが、倍精度のテンソルコアが搭載されていないその他の GPU では提案手法が優位になっていることが分かる。最新の B100 GPU などでは倍精度の性能が大幅に削減されており、今後半導体はますます低精度型の演算に重きを置くようになるため、本手法は多くのアプリケーションで活用されると予想される。

＜引用文献＞

① Qianxiang Ma, Sameer Deshmukh, Rio Yokota, Scalable Linear Time Dense Direct Solver for 3-D Problems Without Trailing Sub-Matrix Dependencies, The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC22), Nov. 2022.

② Qianxiang Ma, Rio Yokota, An Inherently Parallel H^2-ULV Factorization for Solving Dense Linear Systems on GPUs, International Journal of High Performance Computing Applications, accepted, 2024.

③ Hiroyuki Ootomo, Katsuhisa Ozaki, Rio Yokota, DGEMM on Integer Matrix Multiplication Unit, International Journal of High Performance Computing Applications, accepted, 2024.

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Muhammad Ridwan Apriansyah, Rio Yokota	４．巻 48(3)
２．論文標題 Parallel QR Factorization of Block Low-Rank Matrices	５．発行年 2022年
３．雑誌名 ACM Transactions on Mathematical Software	６．最初と最後の頁 1-28
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3538647	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Sameer Deshmukh, Rio Yokota, George Bosilca	４．巻 未定
２．論文標題 Cache Optimization and Performance Modeling of Batched, Small, and Rectangular Matrix Multiplication on Intel, AMD, and Fujitsu Processors	５．発行年 2023年
３．雑誌名 ACM Transactions on Mathematical Software	６．最初と最後の頁 未定
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Hiroyuki Ootomo, Rio Yokota	４．巻 1
２．論文標題 Recovering Single Precision Accuracy from Tensor Cores While Surpassing the FP32 Theoretical Peak Performance	５．発行年 2022年
３．雑誌名 The International Journal of High Performance Computing Application	６．最初と最後の頁 1
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計10件（うち招待講演　0件／うち国際学会　8件）

１．発表者名 Hiroyuki Ootomo, Rio Yokota
２．発表標題 Mixed-Precision Random Projection for RandNLA on Tensor Cores
３．学会等名 Platform for Advanced Scientific Computing (PASC)（国際学会）
４．発表年 2023年

１．発表者名
Qianxiang Ma, Rio Yokota

２．発表標題
O(N) Factorization of Dense Matrices on GPUs Without Trailing Submatrix Dependencies

３．学会等名
SIAM Conference on Computational Science and Engineering (CSE)（国際学会）

４．発表年
2023年

１．発表者名
Muhammad Ridwan Apriansyah, Rio Yokota

２．発表標題
Parallel QR Factorization of Block Low-Rank Matrices

３．学会等名
SIAM Conference on Computational Science and Engineering (CSE)（国際学会）

４．発表年
2023年

１．発表者名
Satoshi Ohshima, Akihiro Ida, Rio Yokota and Ichitaro Yamazaki

２．発表標題
QR Factorization of Block Low-Rank Matrices on Multi-Instance GPU

３．学会等名
The 23rd International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT＇22)（国際学会）

４．発表年
2022年

１．発表者名
Qianxiang Ma, Sameer Deshmukh, Rio Yokota

２．発表標題
Scalable Linear Time Dense Direct Solver for 3-D Problems Without Trailing Sub-Matrix Dependencies

３．学会等名
The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC22)（国際学会）

４．発表年
2022年

１．発表者名
Sameer Deshmukh

２．発表標題
Acceleration of O(N) Solvers for Large Dense Matrices

３．学会等名
Conference on Advance Topics and Auto Tuning in High-Performance Scientific Computing (ATAT2022)（国際学会）

４．発表年
2022年

１．発表者名
Muhammad Ridwan Apriansyah

２．発表標題
Parallel QR Factorization of Block Low-rank Matrices

３．学会等名
Conference on Advance Topics and Auto Tuning in High-Performance Scientific Computing (ATAT2022)（国際学会）

４．発表年
2022年

１．発表者名
Thomas Spendlhofer

２．発表標題
Iterative Refinement with Hierarchical Low-rank Preconditioners Using Mixed Precision

３．学会等名
Conference on Advance Topics and Auto Tuning in High-Performance Scientific Computing (ATAT2022)（国際学会）

４．発表年
2022年

１．発表者名
石井央，横田理央

２．発表標題
深層学習における2次最適化の汎化性能の検証

３．学会等名
第84回情報処理学会全国大会

４．発表年
2022年

１．発表者名 中村秋海，横田理央
２．発表標題 Vision Transformerにおけるバッチサイズの汎化性能への影響
３．学会等名 第84回情報処理学会全国大会
４．発表年 2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	Ｋｈａｎ　Ｅｍｔｉｙａｚ (Khan Emtiyaz) (30858022)	国立研究開発法人理化学研究所・革新知能統合研究センター・チームリーダー (82401)	
研究分担者	大島　聡史 (Ohshima Satoshi) (40570081)	名古屋大学・情報基盤センター・准教授 (13901)	
研究分担者	伊田　明弘 (Ida Akihiro) (80742121)	国立研究開発法人海洋研究開発機構・付加価値情報創生部門(地球情報基盤センター)・副主任研究員 (82706)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関