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　元 年　 ５ 月 ２４ 日現在

機関番号：８２６２６
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００１１６

研究課題名（和文）大規模機械学習のための並列計算基盤の研究

研究課題名（英文）Large-scale machine learning system

研究代表者
中田　秀基（NAKADA, HIDEMOTO）

国立研究開発法人産業技術総合研究所・情報・人間工学領域・研究主幹

研究者番号：８０３５７６３１

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：大規模機械学習システムの構築に向けて、1）大規模分散システムの構築方法、2）大規模化した際の機械学習手法への影響、3）機械学習アプリケーション、の3つのコンテクストで並列して研究を進めた。
1）構築方法に関しては、シミュレータを用いてネットワーク構成と分散実装手法の関連を研究し、比較的プアなネットワークでも十分であることを定量的に示した。2）大規模化した際の機械学習への影響については、独自に新たなシミュレータを開発して検討を行い、学習率の調整が重要であることを示した。3）機械学習アプリケーションとしては、強化学習と画像生成の研究を進め、それぞれ成果を得た。

研究成果の学術的意義や社会的意義
ディープラーニングに代表される機械学習技術が広く普及しつつあるが、これらは大量の計算を伴うため並列分散化して実行することが非常に重要である。われわれは、大規模な並列分散機械学習システムを構成する方法に取り組み、このような計算システムを比較的安価かつ効率的に運用するために必要とされるハードウェアの構成を検討し、比較的安価なネットワークでも十分な性能が得られることを示した。さらに、パラメータを調整することで大規模に並列化しても計算の収束に影響しないように制御することが可能であることを示した。

研究成果の概要（英文）：Aiming at the construction of large-scale machine learning system, we conducted researches in the following three directions; 1) the system architecture in terms of network configuration, 2) effects on convergence, 3) machine learning applications.1) we investigated the relationship between network configuration and distribution method, using existing simulator. We found that relatively poor network configuration suffice the machine learning applications.2) we developed a novel hybrid simulator and investigated the effect, and found that the learning rate is quite important for parallelization. 3) we studied reinforcement learning and image generation.

研究分野：分散計算、 並列計算、機械学習、

キーワード：分散計算　機械学習　ディープラーニング　並列計算

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

インターネットの普及により大規模なデータの蓄積が容易になるとともに、大規模な計算を行う技術が発達したことから、大規模データを大規模な計算機で扱う機械学習技術}が大きく発展している。一方、IoT（Internet of Things）の今後の発展によって、より大規模なデータが日々生成されるようになることが予想される。また Deep Learning に代表される、従来の機械学習よりも遥かに計算量の大きい機械学習アルゴリズムも発達しており、データ量、計算量ともに従来のアプローチでは対応することが難しい。

２．研究の目的

データ量と計算量双方の観点で大規模な機械学習を可能にするシステムを構築することが目的である。このために、以下の 2 つの観点から研究を進めた。

１）　並列分散機械学習に特有な通信パターンを特定し、それに適した計算機アーキテクチャおよびソフトウェア・システムを同定する。

２）　並列化による機械学習の学習速度や精度に対する影響を特定し、それを回避する通信方法、アルゴリズムを一体として提案する。

３．研究の方法

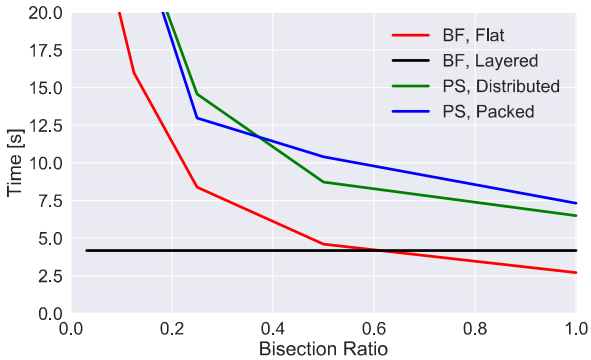
本研究は大規模な計算機環境上でのシステムを対象としているが、つねに大規模な計算機環境を利用することができるわけではなく、仮にできたとしてもパラメータを調整する事はできない。このため、本研究では実際のミドルウェアの開発と並行して、シミュレータを多用した。シミュレータとしては既存の分散システム向けシミュレータである SimGrid を用いるほか、必要に応じて独自に実装した。

４．研究成果

1) 並列機械学習に適したネットワーク構成の研究

データ並列機械学習を実行する計算機におけるネットワークへの要件を明らかにするために、シミュレーションによる定量的評価を行った。並列計算機を導入する場合、ネットワークの費用は全体の費用のうち相当な部分を占める。ネットワークへの要請は対象とする計算の性質に大きく依存する。もし対象となるアプリケーションが帯域を必要としないものであった場合には、ネットワークを比較的プアなものにすることで、大きく予算を削減することができる。われわれは、データ並列機械学習を対象に、いくつかの通信方法を用いた際のネットワーク帯域と通信時間の関係を定量的に評価した。

評価には分散環境シミュレータである SimGrid[1]を用いた。通信方法としては、代表的なパラメータサーバを用いる手法と、サーバを用いずに直接通信する手法を比較した。ネットワークとしては 2 層、3 層のファットツリーネットワークを用いて検証を行った。この結果、機械学習におけるネットワーク負荷は、通常のいわゆる高性能計算における通信と比較して遥かに小さく、したがってネットワークへの要請が小さいことが明らかとなった。右に代表的な結果を示す。横軸はネットワークのリッチさを示す指標で、右側がよりリッチとなる。縦軸は通信にかかる時間を表す。４本のそれぞれ通信方法を表す。うち、３つの手法はネットワークがプアになると速度が低下しているが、１つの手法は影響を全く受けないことが確認できる。この手法を用いれば、ネットワークにかけるコストを大きく低減できる。



本件に関しては、査読付き国際会議 2 件の発表を行っている。現在雑誌投稿の準備中である。

2) データ並列機械学習における耐故障性の研究

大規模な並列計算においては、すべてのノードが健全に動き続けることを期待することはできず、つねにいずれかのノードが予測不可能な形で停止することを前提としてシステムを構成する必要がある。われわれは、SimGrid 上に構築した擬似的な機械学習環境を用いて、さまざまな耐故障アルゴリズムの定量的比較を行った。モデル同期手法の一つであるパラメータサーバを用いた方法では、この方法に固有の耐故障アルゴリズムによって効率的に耐故障性が実現できることを示した。本件に関しては査読付き国際会議発表１件、研究会発表１件を行っている。

3) ハイブリッドシミュレータを用いた、非同期データ並列機械学習の収束性の相関の研究
多くの場合、機械学習のデータ並列化は同期的に行う。つまり定期的に全てのノードで計算を停止し途中経過を交換する。高並列な環境ではこの同期操作のコストが大きくなるため、これを非同期的に行う方法を検討する必要がある。しかし、非同期なデータ並列化による学習過程の収束性への影響は明らかではない。われわれは、通信遅延を任意に挿入でき、なおかつ実際に機械学習を行うハイブリッドシミュレータを構成、実装した。このシミュレータは Python のコルーチンを用いて構成されており、大きな性能低下なく、実アプリケーションを動作させる事が可能である。アプリケーションを動作させた結果、小規模な環境では学習率設定が、実際には保持していないような規模の大規模環境では学習を不安定にすることを確認し、これを通じてシミュレータの有効性を確認した。研究会発表 2 件を行っている。

4) 並列機械学習フレームワークのスーパーコンピュータへの適用
いくつかの並列機械学習システムが提案されているが、比較的小規模な環境での実行を前提としている場合が多く、既存のいわゆるスーパーコンピュータ環境では動作しない場合がある。例えば UCB で開発された Ray[2]は、起動機構と通信機構の 2 点でスーパーコンピュータ環境との相性が悪く動作が難しい。われわれは Ray を一つのサンプルとして、並列機械学習システムのスーパーコンピュータ環境への適用手法の研究を行った。具体的にはネットワーク通信レイヤとして並列計算機環境ではデファクトスタンダードとなっている MPI を利用するよう改変するとともに、バッチキューイングシステムからのジョブ対応を可能とした。研究会発表 2 件を行っている。

5) 大規模データ処理基盤の改良
大容量データ入力においては、従来の分散ファイルシステムでは多数の計算機からのランダムに近いデータ要求に対して十分な性能を確保することができないため、メモリ上に確保されたキャッシュを効率的に用いることが鍵となる。本課題では並列分散システム Spark の中間データキャッシュ機構に着目し改良を進めた。具体的にはキャッシュのフラッシュアルゴリズムを改良し、不必要なデータコピーが生じないようにした。この結果大幅な速度向上を得ることができた。査読付き国際ワークショップ 1 件を行っている。

参考文献
[1] Henri Casanova, Arnaud Giersch, Arnaud Legrand, Martin Quin- son, and Fre´de´ric Suter. Versatile, scalable, and accurate simulation of distributed applications and platforms. Journal of Parallel and Distributed Computing, 74(10):2899-2917, June 2014.

[2] Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Lian. Ray: A Distributed Framework for Emerging AI Applications,USENIX OSDI, 2018.

５．主な発表論文等

〔雑誌論文〕（計　1　件）

A Quantitative Analysis on Required Network Bandwidth for Large-Scale Parallel Machine Learning, Mingxi Li, Yusuke Tanimura, Hidemoto Nakada , MOD 2017 - (The Third International Conference on Machine Learning, Optimization and Big Data) , LNCS vol.10710 , pp. 389-400 , 2017

〔学会発表〕（計 10 件）

How Much Should We Invest for Network Facility: Quantitative Analysis on Network 'Fatness' and Machine Learning Performance, Duo Zhang, Mingxi LI, Yusuke Tanimura, Hidemoto Nakada , Workshop on ML Systems in NIPS 2017 , 2017

Understanding and Improving Disk-based Intermediate Data Caching in Spark, Kaihui Zhang, Yusuke Tanimura, Hidemoto Nakada, Hirotaka Ogawa , Scalable Cloud Data Management Workshop 2017 in IEEE BigData , pp. 2426-2435 , 2017

Adaptation of Ray, a distributed framework for machine learning, to MPI-based environment, Tianlun WANG, Yusuke Tanimura, Hidemoto Nakada , 信学技法 IEICE-CPSY , 2018

A Hybrid Simulator to Analyze Gradient Staleness Effect, Duo Zhang, Yusuke Tanimura, Hidemoto Nakada, 情報処理学会 システムソフトウェアとオペレーティング・システム研究会

Asynchronous Deep Learning Test-bed to Analyze Gradient Staleness Effect, Duo Zhang, Yusuke Tanimura, Hidemoto Nakada , 信学技法 IEICE-CPSY , 2018

A Quantitative Analysis of Fault Tolerance Mechanisms for Parallel Machine Learning Systems with Parameter Servers, Mingxi Li, Yuusuke Tanimura, Hidemoto Nakada , ACM IMCOM 2017 , 2017

A Performance Evaluation of Distributed TensorFlow,
Tianlun Wang, Yusuke Tanimura, Hirotaka Ogawa, Hidemoto Nakada , 研究報告ハイパフォーマンス・コンピューティング（HPC）2017-HPC-161 , pp. 1--6 , 2017
A study on Network Structure and Parameter Exchange Method in large-scale Cluster for Machine Learning, Dou Zhang, Rei Mingxi, Yusuke Tanimura, Hidemoto Nakada , 信学技報, vol. 117, no. 153, CPSY2017-29 , pp. 145-150 , 2017

大規模機械学習向けクラスタにおけるネットワーク構造とパラメータ交換手法,
黎 明曦, 谷村 勇輔, 中田 秀基 , cross-disciplinary Workshop on Computing Systems, Infrastructures, and Programming , 2017

Spark におけるディスクを用いた RDD キャッシングの高速化と 効果的な利用に関する検討,
張 凱輝, 谷村 勇輔, 中田 秀基, 小川 宏高 , cross-disciplinary Workshop on Computing Systems, Infrastructures, and Programming , 2017

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）

○取得状況（計　0　件）

〔その他〕
ホームページ等

https://sites.google.com/site/infrawarelab/
https://sites.google.com/site/hidemotonakada/

６．研究組織

(1)研究分担者
N/A
(2)研究協力者
N/A