（ 続紙 １ ）

京都大学	博士（情報学）	氏名	小山田 創哲
論文題目	Toward Scalable Reinforcement Learning via Massive Batching （大規模バッチングによるスケーラブルな強化学習の実現に向けて）		

（論文内容の要旨）

エージェントが環境と相互作用しながら学習を行う強化学習は、単なる人間の模倣にとどまらず、人間を超える能力を持つ人工知能（AI）を実現する可能性を秘めている。しかし、現在の強化学習アルゴリズムは、一般的に、膨大な学習データを必要とし、また学習が遅く非効率的である。本論文では、「計算をレバレッジする」アプローチ、特に大規模なバッチ処理とGPUアクセラレータの活用に基づき、この問題の解決を目指したものである。本論文は以下の６章から構成される。

第１章は序論であり、強化学習の「父」R. Suttonの言葉を引用しながら、本論文のモチベーションとなる「計算をレバレッジする」ことのAI研究における役割の重要性について振り返っている。第２章は予備知識として、深層強化学習を概観した上で、深層強化学習のボトルネックが実際には、勾配計算やパラメータの更新ではなく、環境とエージェントの相互作用によるデータ生成のパートにあることが一般的であることを確認している。そしてボトルネックであるデータ生成を並列化する手法を、MIMD/SIMDの概念を用いて２つに大別している。本論文でフォーカスする後者のGPUアクセラレータを活用したベクトル化による並列化手法の利点を強調しながらも、逆にその適用範囲の限界についても議論している。

第３章では、連続状態空間タスクで有効なGPU上でのベクトル化された環境を使った並列化手法が、複雑な条件分岐を伴う離散状態空間の環境においても有効かどうかを検証している。バックギャモン、チェス、将棋、囲碁などAI研究における重要なベンチマークを含む20以上のゲームがGPU上で動作する環境群、Pgxを実装することで、バッチ処理に基づくフルGPU実装の有効性を構成的に確認した。結果として、Pgxは既存の公開されている実装より、単一GPUで10倍、8GPUで100倍以上としてスループットが改善されることを示した。また、PPOやAlphaZeroといった主要な強化学習アルゴリズムもこのPgx上で高速に動作することを検証した。

第４章では、Pgxによる大規模バッチ化が実際に強化学習研究を加速できることを実証するため、マルチエージェント不完全情報ゲームの一つであるブリッジビッディングAIを構築した。結果として、PgxによるGPU上での大規模なバッチ処理と、マルチエージェント環境で学習を安定させる既存の学習手法のシンプルな組み合わせだけで、ブリッジビッディングAIのベンチマークにおいて先行研究を凌駕するスコアを達成できることを示した。

第５章では、大規模ベクトル化・バッチ処理による別の側面について考察をしている。GPUを活用した大規模なバッチ処理によれば、計算効率が向上できる一方で、学習器へのフィードバックが遅れるため、逐次的な処理に比べて学習性能が低下する可能性がある。この問題について、強化学習の最もシンプルな形式である確率的バンディットでの純探索問題を用いて考察している。Sequential Halving (SH; Karnin et al., 2013) はシンプルかつ効率的な純探索問題アルゴリズムとして知られているが、本論文では、このSHを自然にバッチ拡張した Advance-first Sequential Halving (ASH) は、現実的な条件のもとで性能劣化を全く引き起こさないことを理論的・実験的に示した。こうしたSHのバッチ化に対する頑健性を踏まえ、モンテカルロ木探索においても、ルートノード上でSHを用いることで複雑なゲームツリーに対する探索におい

ても性能低下を抑制しつつ、バッチ化により計算効率を向上させることができることを示した。

第６章は結論であり、本論文で得られた成果をまとめている。即ち、３章ではGPUアクセラレータを用いた環境のバッチ化並列による強化学習の適用範囲が考えられているよりも広いことを示し、４章ではその高い有効性をデモンストレーションした。５章ではそうしたバッチ化並列環境に適したアルゴリズムの実例を示し、今後の強化学習研究において、バッチ化並列された環境を前提として、その性質を有効活用するアルゴリズムを開発・設計することの重要性を示唆している。

（続紙 ２ ）

（論文審査の結果の要旨）

本論文は、強化学習における典型的なボトルネックである環境とエージェントの相互作用によるサンプル生成において、GPUアクセラレータを用いた並列化手法の適用範囲と有効性に関して論じた論文である。本論文で得られた主要な結果及び知見は以下の通りである。

（１）連続状態空間環境やトイ・プロブレムでは、状態遷移の行列演算をGPU上で実装することで、CPUとGPU間のデータ転送コストをかけずに、GPU上で大規模なバッチ処理を実現するアプローチが可能である。しかしながら、チェスや囲碁のようなAI研究における重要なベンチマークである離散状態空間環境においては、複雑な条件分岐を伴うため、こうしたGPUを活用したバッチでの状態遷移が有効かどうかは明確ではなかった。本論文ではチェスや囲碁を含む20以上のGPU上で動作するゲーム環境群、Pgxを実装することで、複雑な離散環境でもGPUを活用したバッチ並列環境が有効であることを構成的に示した。

（２）ブリッジビッディングAIにおいて、先述のGPUを活用したバッチ並列環境による強化学習と、マルチエージェント強化学習におけるシンプルな既存技術の組み合わせによって、ベンチマークAIに対する対戦成績において先行研究を凌駕することができることを示した。このGPUを活用したバッチ並列環境による強化学習が実際に強化学習研究を加速させることのデモンストレーションによって、その有効性を示した。

（３）大規模なバッチ化は計算効率面での優位性があるが、遅延フィードバックによる性能劣化を引き起こす可能性がある。強化学習の最もシンプルな形である多腕バンディットの純探索問題において、Sequential Halving (SH; Karnin, et al., 2013) アルゴリズムがそうした性能劣化に非常に頑健であることを理論的・実験的に示した。また、モンテカルロ木探索へも応用することで強化学習におけるバッチでのプランニングアルゴリズムの計算効率向上に貢献できることを示した。

まとめるに、本論文は、GPU上で動作するように大規模にバッチ化する手法が、連続状態空間の環境だけでなく、囲碁のように複雑な条件分岐を含む離散状態空間の環境にも応用可能であることを示した。また、ブリッジビッディングAIにおいて先行研究を凌駕する性能を達成することで、このアプローチが強化学習研究を実際に加速することを示した。バッチ処理に伴うフィードバック遅延によるパフォーマンス低下の懸念についても議論を行い、適切なアルゴリズムの選択により性能劣化を避けられることを示した。これらの成果は、今後の強化学習研究において、大規模なバッチ処理を念頭に置いた強化学習アルゴリズムの設計の必要性を示している。また、比較的チープな計算機環境を有する一般の研究者にも利用可能な強化学習テストベッドを提供することで、分野全体の底上げにも寄与しえるものである。

よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、令和6年10月17日、論文内容とそれに関連した事項について試問を行った結果、合格と認めた。また、本論文のインターネットでの全文公表についても支障がないことを確認した。