修士学位論文要約（令和４年３月）

確率的状態遷移を伴う深層強化学習のエッジコンピューティング向けハードウェアに関する研究

小口　大輔

指導教員：佐藤　茂雄，　研究指導教員：山本　英明

Study on Hardware Implementation of Deep Reinforcement Learning with Probabilistic State Transitions for Edge Computing Applications

Daisuke OGUCHI

Supervisor: Shigeo SATO,　Research Advisor: Hideaki YAMAMOTO

Reinforcement learning is promising as a machine learning paradigm in edge computing. However, reinforcement learning is computationally expensive, and the development of dedicated circuits is essential when implementing it in devices with limited circuit resources and power consumption. In this study, we developed dedicated circuits for the Q-learning algorithm and deep Q-learning using RTL design and investigated the relationship between the bit length of floating-point operations and the learning performance of the reinforcement learning algorithm. We found that when solving the FrozenLake maze problem with Q-learning, the learning performance of 16-bit floating-point operations is comparable to that of 64-bit CPU operations. It was also found that an approximate implementation of the Boltzmann action selection method is possible while reducing the cost of the exponential function. In addition, we developed a prototype dedicated circuit for deep Q-leaning, in which action selection is implemented on an FPGA and the action value function is approximated by a deep neural network running on a PC. Our results provide practical guidelines for designing dedicated reinforcement learning hardware with minimal circuit resources and power consumption.

1. はじめに

IoT 化の進展に伴い，よりパーソナライズされた情報システムの実現が期待されている．その中で AI に代表される高度な情報処理をエッジデバイス上に実装するためのハードウェア技術と，報酬を用いて学習を進めていく強化学習アルゴリズムを組み合わせたシステムの実装に注目が集まっている．しかし強化学習アルゴリズムは計算コストが高く，エッジデバイス上に実装する際には専用ハードウェアの開発が不可欠である．

本研究では，エッジコンピューティングに向けた強化学習専用ハードウェアの設計と実装を目的する．状態遷移が確率的な環境モデルにおいて，Q 学習を専用で処理する回路をレジスタ転送レベル（RTL）で設計し，回路リソースの削減と学習性能の関係について調査する．さらに，Q 学習を深層化した Deep Q-learning[1]に置き換え，深層強化学習の専用ハードウェア実装時の課題について議論する．

2. モデルフリー強化学習による環境探索

強化学習では，教師あり学習とは異なり明確な教師データは与えられず，代わりに行動の選択肢と一連の行動に対する報酬が与えられる．本研究で用いた基礎的な強化学習の一つである Q 学習アルゴリズムを定式化すると以下の通りである．

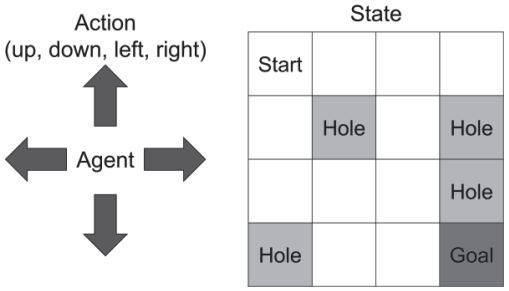

図 1　FrozenLake 迷路課題における 4 つの行動（左）と 16 つの状態（右）

$$Q(s,a) = R(s,a) + \gamma \max_{a'} E[Q(s',a')] \quad (1)$$

状態sで行動aを選択した際の報酬の期待値を行動価値$Q(s,a)$と呼ぶ．この行動価値の更新式は，

$$target = R(s,a) + \gamma \max_{a'} E[Q(s',a')] \quad (2)$$

$$loss = target - Q(s,a) \quad (3)$$

$$Q(s,a) \leftarrow Q(s,a) + \alpha \times loss \quad (4)$$

で表される．エージェントは基本的に行動価値が一番高い行動を選択するが，学習する際は行動方策に従いし，未知の状態を探索する必要がある．行動

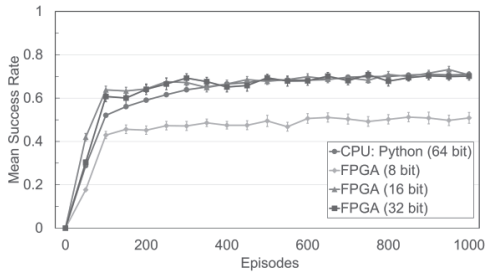

図 2　数値表現の bit 幅削減と学習性能の関係．エラーバーは標準誤差を示す（FPGA：n=30）．

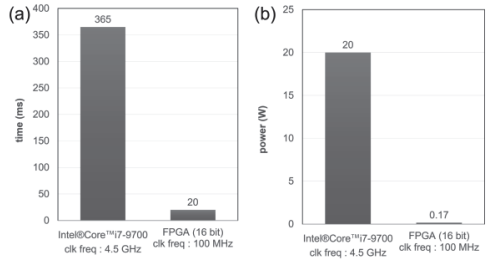

図 3　ε-greedy 手法を実装した場合の(a)処理時間，(b)消費電力の比較

方策の代表的なものとして，ε-greedy 手法と Boltzmann 行動選択手法が挙げられる．Boltzmann 行動選択手法は確率的な行動選択が可能であるが，指数関数の計算を含むため，回路リソースを削減するための工夫が必要である．

また，実環境のような連続的な状態空間を考える場合には，価値関数や方策関数をニューラルネットワークで近似した深層強化学習（DRL）が非常に有効である．そのうち Deep Q-learning（DQL）は，Q 学習アルゴリズムの行動価値関数をニューラルネットワークで近似したものである．

3. FPGA で動作可能な Q 学習システムの実装

本論文では，Verilog-HDL を用いて Q 学習アルゴリズムの処理を RTL で記述し，設計した回路の動作シミュレーションと FPGA 実装を行った[2]．また，実装した強化学習システムの性能検証用タスクとして，状態遷移が確率的である FrozenLake 迷路課題を採用した（図 1）．学習時の数値表現の bit 幅削減とエージェントの学習性能の関係について調査した結果を図 2 に示す．32 bit および 16 bit 浮動小数点表記の場合は成功率が約 70%に到達し，これは CPU 計算（64 bit）と同程度であった．論文中では，Boltzmann 行動選択手法の近似実装を行い，その結果についても述べている．図 3 は，設計した Q 学習専用回路を FPGA に実装した場合の処理時間と消費電力の見積もりを示している．専用回路の設計に

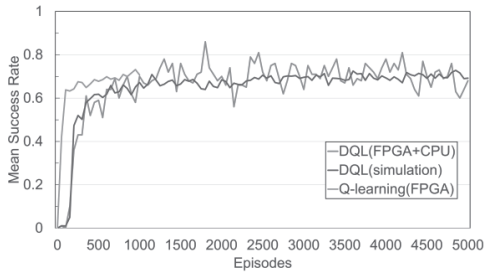

図 4　実装した DQL システム（FPGA+PC，サンプル数 n=2）とシミュレーションの成功率の比較

より，CPU 計算に比べて処理時間と消費電力が大幅に削減できることを明らかにした．

4. 深層強化学習のハードウェア化とエッジコンピューティング応用

DQL の専用回路に関して，シミュレーションと実機への実装を行った．先に実装した Q 学習アルゴリズム専用回路の行動価値関数を深層ニューラルネットワーク（DNN）に置き換え，その学習性能の調査を行った．尚，実装までの時間の都合上，本論文では DNN の処理を PC で行っている．図 4 は DQL システムの学習性能の結果である．成功率は約 70%であり，これは CPU 計算と同程度であった．今後は，DNN を専用で処理するハードウェアと連携したシステムの実装を考えている．

5. まとめ

本研究では，強化学習を専用で処理するハードエアを実装するため，Q 学習アルゴリズム専用回路の RTL 設計を行い，シミュレーションと FPGA 実装によりその学習性能を評価した．処理フローの RTL 設計により，学習に要する処理時間と消費電力が CPU 計算に比べて大幅に削減されることを示した．さらに，内部の浮動小数点表記の数値精度と学習性能の関係を調査し，bit 幅を 16 bit にまで制限した場合でも学習性能が維持されることを明らかにした．また，Boltzmann 行動選択手法の近似実装を行い，指数関数の計算コストの削減と学習性能の維持が可能であることを示した．加えて，Q 学習を深層化した DQL を用いたシステムを実装し，その動作確認と学習性能の検証を行った．本結果は，確率的な状態遷移を伴う実環境での動作を想定した強化学習専用回路を実装する上で重要な知見である．

文献

1) V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, *et al*., *Nature*, **518** (2015) 529-533.
2) D. Oguchi, S. Moriya, H. Yamamoto and S. Sato, *NOLTA, IEICE* (in press).