技術報告 令和 3 年度科学研究費補助金（奨励研究）
深層強化学習による時系列処理学習
～リザバを使った新しい経験リプレイの提案～

大分大学理工学部技術部

松木俊貴

1．研究の目的

多層のニューラルネットワーク（NN）をデータに基づいて学習させる深層学習は，既存の手法では難しかったさまざまな情報処理が可能であることが示されたことで注目を集めている．さらに，エージェント自身が探索して得られた経験に基づいて学習を行う強化学習と深層学習とを組み合わせた深層強化学習が，明示的な教示を与えることが難しいゲームや制御などのタスクに対して有効な技術として盛んに研究が行われている．深層強化学習では，NN が環境に対し行動選択をし，報酬と状態を受取る相互作用の中で学習を行う．また，学習中の各時刻での状態・行動・報酬・行動後の状態といった「経験」を記録しておき，それらをランダムにサンプリングして学習することで経験の偏りを防ぐ「経験リプレイ」という手法が広く用いられる．

以前の状況がしばらく後になって影響を及ぼすような環境の場合，NN は過去からの文脈を考慮して行動選択する必要がある．そのような環境では，再帰構造を持つリカレント NN（RNN）を用いた強化学習が行われる．しかしその場合，文脈を無視してある時刻の瞬間的経験だけを独立にサンプリングしてもうまく学習できない．そこで，様々な経験リプレイ手法が研究されているが，学習に必要なメモリ量や計算量が増大するという課題がある．そこで本研究では，シンプルな手法によって時系列処理を学習できるリザバーネットワーク（RN）と呼ばれる特殊な RNN を使った経験リプレイを導入し，それにより時系列処理が必要な制御タスクの学習が可能であること，そして RN の出力層に通常用いられる単層の線形層ではなく多層の NN を用いることで学習性能が向上することを示した．

2．研究方法

本研究では，タスク環境から得られた観測o_tをそのままメモリに保存するのではなく，RN の出力を同時に保存する手法を用いた（図 1）．RN は内部の結合構造を学習せず，特殊な初期値で固定する．そして RN は，入力を時空間パターンに展開し，過去の入力の記憶を保持しながら非線形処理した出力を生成する一種のフィルターのような役割を担う．この出力を経験として保存し学習に用いることで，計算量の増大や学習の不安定性をもたらす「時間を遡る処理」を伴わずに時系列処理を強化学習することが可能となる．また本研究では，RN において唯一重み値の更新が行われるリードアウト（出力層）として，通常用いられる単層の線形フィルターではなく多層の NN を用いた．また，学習には深層強化学習の手法として広く知られている Deep Q-Network(DQN)の手法を用いた．その際，学習性能の向上に寄与することが報告されている Double DQN や Reward Clipping の技術を導入した．

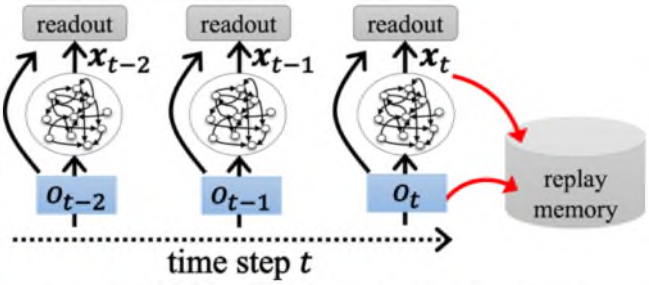

図 1　RN を使った経験リプレイ．環境から得られた観測o_tとそれを RN に入力して得られた出力をリプレイメモリに保存する．

時系列処理の学習性能を検証するため，非営利団体 Open AI が提供し，強化学習のベンチマークとして広く用いられているライブラリ Gym に含まれる 4 つの古典的制御タスク Cart Pole, Mountain Car, Acrobot, Pendulum を，エージェントが速度や角速度といった情報を得られないように改造して用いた．また，環境から観測される入力は-1 から 1 の範囲に正規化して RN とリードアウトに与えた．学習成功の条件はエージェントが各タスクを 10 試行連続でクリアすることとした．

3．検証結果

図 2 にそれぞれのタスクの学習曲線を示す．Cart Pole タスクはカートに取り付けられた振り子を，カートを適切に動かすことによって最大 200 ステップの間振り上げ続けることを目的とする．図 2(a)から，試行を重ねるにつれて振り子を立たせ続けられるステップ数が上昇していることがわかる．そして，第 178 試行目においてエージェントは 10 試行連続で条件を達成することに成功した．Mountain Car タスクは，上限 200 ステップが経過する前に，一度バックして後方の山に登りそこから斜面を下る時の勢いを利用して前方の山の上のゴールに到達することを目的とする．図 2(b)から，第 140 試行前後から上限ステップに到達する前にゴールできるケースが現れ始めている様子がわかる．ここでは，第 223 試行においてエージェントは 10 試行連続で条件を達成することに成功した．Acrobot タスクは，二重振り子に適切な向きのトルクを加えて先端を規定の高さ以上に振り上げることを目的とする．図 2(c)より，試行回数の進行とともに上限の 200 ステップ以内で振り子の振り上げに成功するケースが増え始めていることがわかる．また，第 117 試行においてエージェントは 10 連続でタスクに成功した．Pendulum タスクは，振り子に適切な向きのトルクを加えて上向きに振り上げたあと，その状態を維持することを目的とする．図 2(d)から，試行回数が進むごとに総獲得報酬が上昇しており，エージェントが適切に学習しているこ

とがわかる．以上の結果は，RN の出力を経験としてメモリに保存する手法により，時系列処理が必要な古典的制御タスクの強化学習が可能であることを示している．

次に，多層のリードアウトを用いた場合と，通常よく用いられる単層のリードアウトを用いた場合の性能比較を行った．RN のパラメーターのうち学習性能に最も影響を与えるものの一つに相互結合重み値行列のスペクトル半径を決定するgがある．一般にgが大きいと長期間記憶を保持できるが RN の性能を発揮するために通常$g < 1$とすることが求められる．一方で入力との関係によって$g \geq 1$の時に RN の性能がうまく引き出されることも多い．タスクに応じて適切なgの範囲は異なることから，ここではgを0から2まで変化させ，異なる乱数シードの元で 100 回の学習を実施し，そのうち学習に成功した割合を調べた．その結果を図 3 に示す．図から，いずれのタスクにおいても単層のリードアウトを用いた場合の方が，学習できないかもしくは成功確率が低い傾向にあり，リードアウトを多層にすると学習性能が向上していることが確認できる．この結果は，RN を用いた深層強化学習において，多層のリードアウトを用いることが有効であることを示唆している．

4．まとめ

本研究では，RN を用いた経験リプレイを導入した深層強化学習により，時系列処理が必要な古典的制御タスクの学習が可能であること，および多層のリードアウトを用いることでその学習性能が向上することを示した．今後は今回用いた DQN 以外の強化学習手法での検証や，RN に不向きな画像などの高次元入力を扱うタスクの学習方法について検討する必要がある．

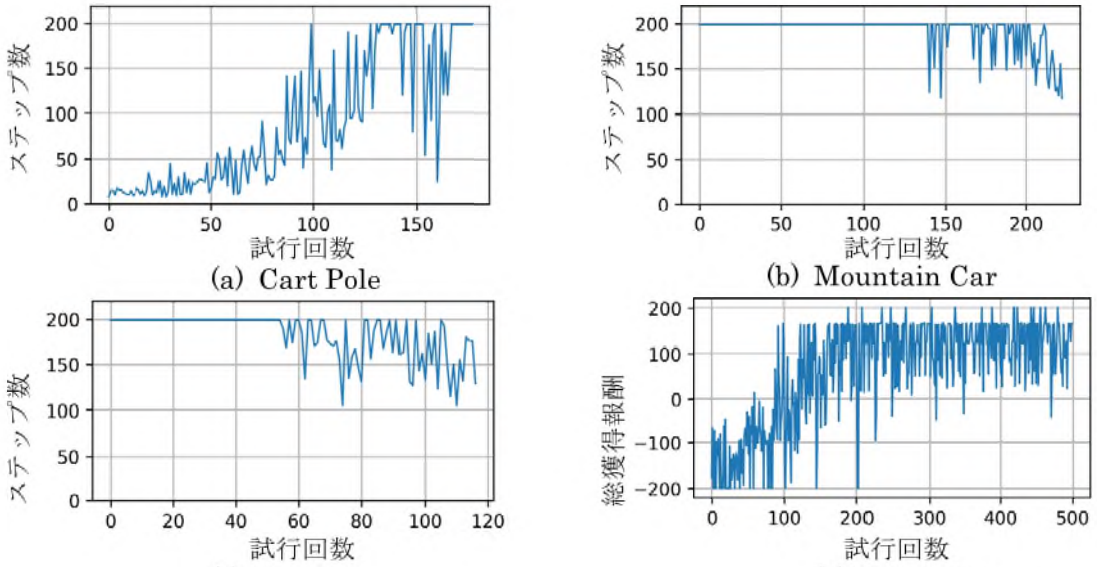

図 2　各タスクにおける学習曲線．横軸は学習の試行回数を示しており，(a)-(c) の縦軸は各試行におけるステップ数，(d)の縦軸は各試行での総獲得報酬値を示している．

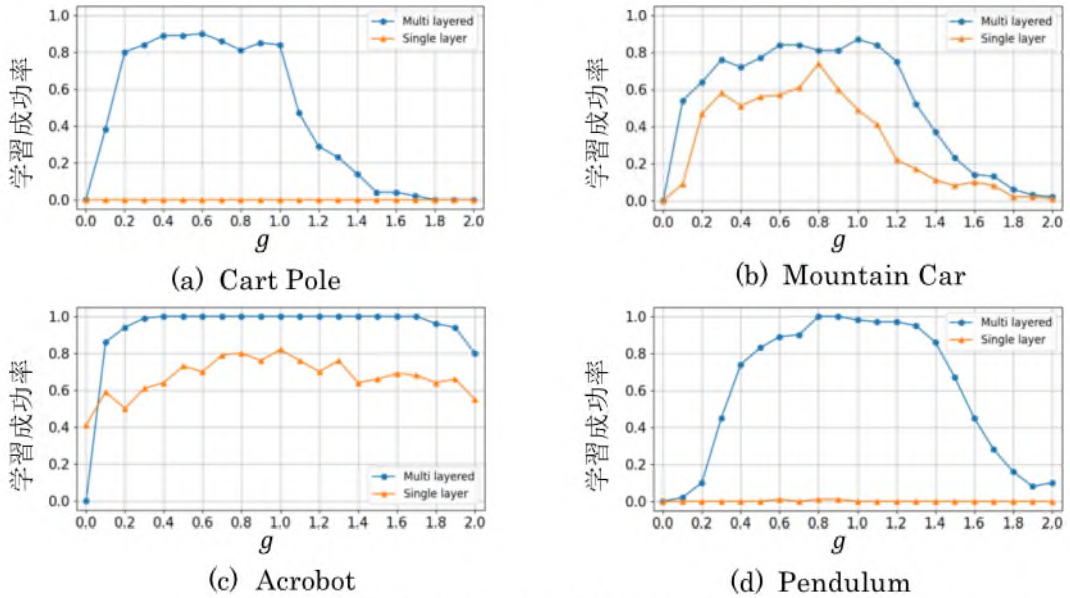

図 3　RN の相互結合重み値のスペクトル半径gと学習成功率.青線は多層リードアウト，オレンジは単層リードアウトの結果をそれぞれ示している．