技術報告　「科学研究費助成事業(奨励研究)」
深層強化学習で非線形な制御が学習できるか
～物理演算ゲームの学習を通じた検証～

○松木俊貴
大分大学理工学部技術部

1.　研究背景・目的

近年，あらゆる分野において深層学習 (Deep Learning) が既存のアプローチを凌駕することが示され大きな注目を集めている[1]．このことは，柔軟かつ並列な Neural Network (NN) が，大量のデータをもとにして学習することで，人の手で設計されたシステムよりも優れた性能を獲得できることを示している．また，Atari 社の TV ゲームにおいて人間と同等あるいはそれ以上のハイスコアを記録した Deep Q-Network の成功をきっかけとして，明示的な教示なしに，探索により得られた報酬と罰のみを通じて学習を行う強化学習と深層学習を組み合わせた深層強化学習の研究が盛んに行われている[2]．囲碁のプロ棋士に勝利した AlphaGo は，この手法を取り入れた学習により，人の手では設計できないレベルの性能を自律的に獲得することに成功した．

将来，我々の生活の中でロボットが活躍することが期待されている．工場のような閉じた環境ではなく，実社会の中で人間と協力して働くことのできるようなロボットの実現のためには，事前には想定しきれないどのような状況下でも適応する能力を持ち，画像を始めとするあらゆる情報を含んだ高次元のセンサ情報をうまく統合し，与えられた目的を達成するために非線形に動きをかえていけるような制御が必要だと考える．だがこれは，事前に与えた目標起動，対象モデル，ゲイン調整などに基づく従来型の制御による実現は難しい．しかし，高次元な情報を非線形に統合し処理することに長けている NN と探索と報酬に基づき自律的に学習を行うことのできる強化学習とを組み合わせた深層強化学習であれば，このような制御能力の獲得も可能ではないかと期待される．

本研究では，深層強化学習を非線形な制御が求められるタスクに適用することで，人の手による設計を離れ，自律的に合目的的な制御則を獲得できる枠組みが実現できないか検証を行う．

2.　研究方法

本研究では Roll-A-Ball ゲームを制御タスクとして用いる．このタスク環境では，ボードの上に玉，障害物，落とし穴，ゴールなどが存在し，エージェントはボードの傾斜を傾けることにより，玉を転がしてゴールさせることを目的とする．このタスクでは，ボード上のオブジェクトの位置はゲーム開始ごとにランダムに変わるため，事前に目標軌道を定めることは難しく，障害物を避けながら非線形に傾斜角を変えていかなければならない．

学習には Actor-Critic と呼ばれる強化学習手法を用いて行う．ある時刻tにおいて，ネットワークは環境から得られたセンサ信号を入力とし，現在のエージェントの状態価値を意味する Critic 信号C_tと，ボード傾斜角を表す Actor 信号$\boldsymbol{A}_t$を出力する．出力された Actor 信号に探索成分として乱数ベクトル$\boldsymbol{rnd}_t$を加え，その値により環境内に存在するボードの傾斜角を変更する．エージェントは出力に乱数が加えられることで探索を行い，落とし穴に落ちれば-0.8の罰が，ゴールする

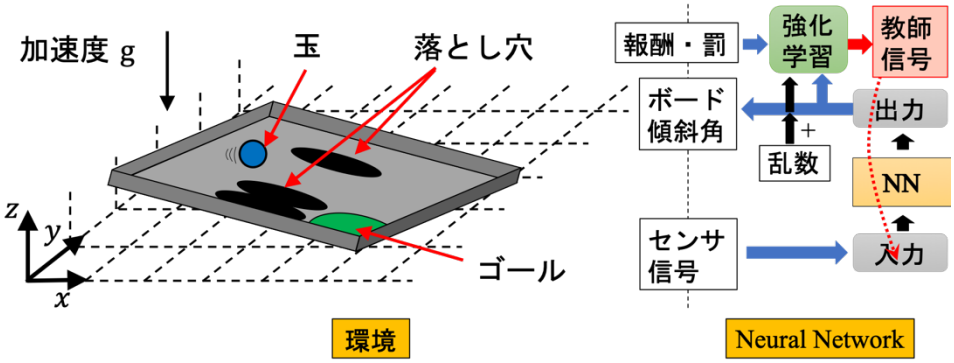

図1 全体構成

ことができれば1.0の報酬がr_tとして与えられる．一つ前の時刻 $t-1$での Critic 信号の教師信号は次式のように与えられる．

$$V_{t-1}^{teach} = V_{t-1} + \hat{r}_{t-1} = r_t + \gamma V_t \tag{1}$$

ここで，$\hat{r}_{t-1}$は時刻 $t-1$における TD 誤差であり，次式により得られる．

$$\hat{r}_{t-1} = r_t + \gamma V_t - V_{t-1} \tag{2}$$

$\gamma = 0.99$は割引率である．Actor 信号の教師信号は次式により得られる．

$$\boldsymbol{A}_{t-1}^{teach} = \boldsymbol{A}_t + \hat{r}_{t-1}\boldsymbol{rnd}_{t-1} \tag{3}$$

式(1)(3)により得られた教師信号をもとに，誤差逆伝播法により NN の学習を行う．

学習主体となるエージェントは時系列データを処理することができるリカレントニューラルネットワーク(RNN)で構成される．RNN は学習のために時間を遡る処理を必要とすることから，一般的に学習の収束が遅く，不安定であるという課題がある．そのような課題を回避する一つの手法として，リザバコンピューティングと呼ばれる特殊な中間層を持つRNNを用いるアプローチが提案されている[3]．リザバコンピューティングでは内部の重み値を固定し，ダイナミクスを抽出する読み出しユニットのみを学習する．我々のグループは，このリザバと多層の NN を組み合わせたアプローチにより時間を遡る処理なしに記憶タスクを強化学習できることを示した[4]．本研究においても同様のネットワークを用いる．ネットワークの構成を図 2 に示す．入力層の次にリザバ層があり，その上層に通常の多層ニューラルネットを重ねる形で構成されている．赤い矢印で示した部分の結合重み値のみを学習していく．

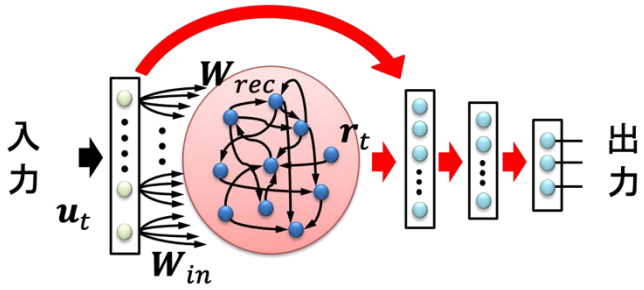

図 2 エージェントネットワーク

ニューロンを$N = 200$個もつリザバ内部の内部状態はベクトル$\boldsymbol{x}_t \in \mathbb{R}^N$で表され次式により与えられる．

$$\boldsymbol{x}_t = \lambda \boldsymbol{W} \quad \boldsymbol{r}_{t-1} + \boldsymbol{W}_{in}\boldsymbol{u}_t \tag{4}$$

ここで，$\boldsymbol{W}_{rec}$は10%の割合で疎結合しているリザバ層の相互結合重み値行列であり，一様乱数により値を生成した後，自身のスペクトル半径で正規化することで固有値の大きさの上限が 1 になるように決定する．$\lambda = 0.95$は$\boldsymbol{W}_{rec}$のスケールを決定するパラメータである．$\boldsymbol{W}_{in}$は入力層からリザバ層への結合重み値で，-0.5から0.5の一様乱数で与えられる．$\boldsymbol{r}_t$はリザバニューロンの出力で，$\boldsymbol{x}_t$を双曲線正接関数に代入することで得られる．$\boldsymbol{u}_t$は環境からネットワークが得られる情報の入力ベクトルである．このように構成されたリザバネットワークは与えられた入力の情報を内部のダイナミクスの中に取り込み長期間保持することができる．リザバ内に保持された情報を抽出するリードアウトユニットだけを学習するため時系列データ処理をシンプルに実現することができる．従来リードアウトユニットは層構造を持たないユニットで構成されるが，本研究では多層の NN を用いた．

図 3 に学習の全体構成を示す．環境には二つの固定された落とし穴とランダムに出現する落とし穴を配置する．

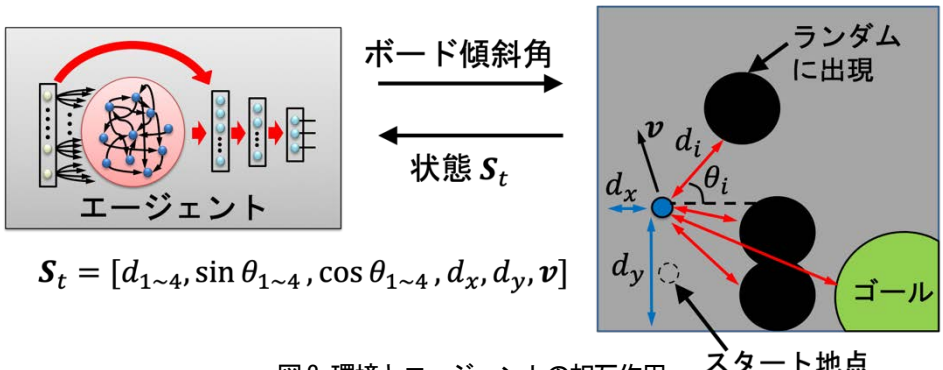

図 3 環境とエージェントの相互作用

毎試行開始時，玉はスタート地点に置かれており，エージェントはボードを傾けることで，転がる玉をゴールエリアへと導く．エージェントはボールと各落とし穴及びゴールエリアとの距離・相対角度，ボールのボード上での位置・速度の情報を環境から得ることができ，それらの情報をもとにボードの傾斜角と状態価値を出力する．

3. 結果

エージェントに対し 35000 試行学習を行わせた時の学習曲線を図 4 に示す．横軸は試行回数，縦軸は 1000 試行ごとにゴールすることができた回数の割合を示している．この結果から，学習試行が進むにつれてエージェントが玉をゴールに導くことができるように学習できていることがわかる．

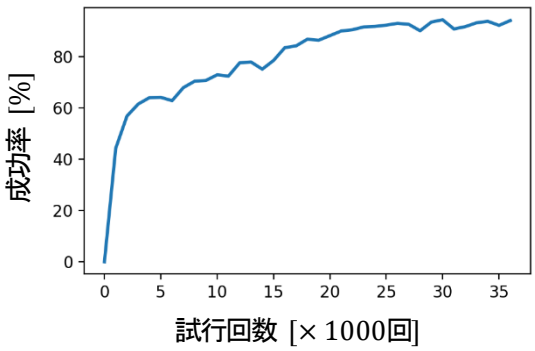

図 4 エージェントの軌道

図 5 に学習初期と学習終了後の転がる玉が描く軌道を赤い線で示す．学習初期では乱数ベクトルにより探索が行われ，玉がボード上を転がりまわっていることがわかる．学習後の軌道では，エージェントが学習したことでランダムな位置にある落とし穴をうまく避けながら玉をゴールへと導くことに成功していることがわかる．

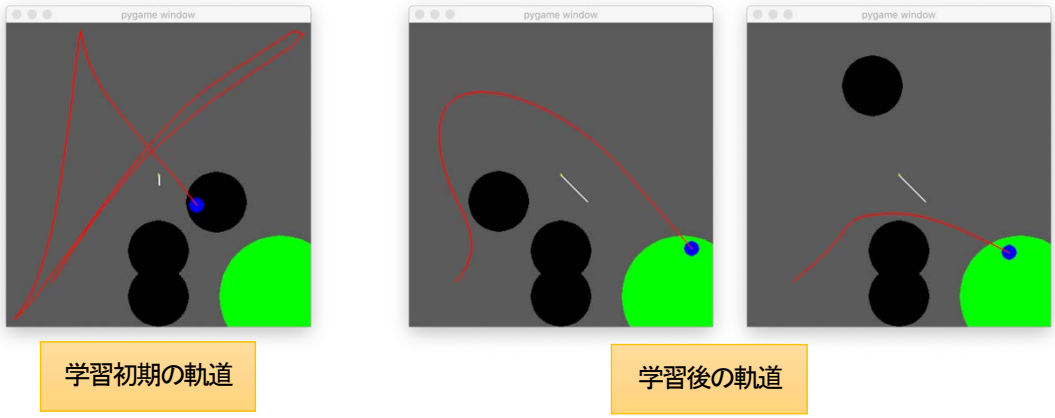

図 5 エージェントの軌道

4. 今後の課題

計画段階では，エージェントに環境の画像を直接与えて学習を行わせる予定であった．しかし，リザバネットワークに直接画像のような高次元のデータを与えることはできないため，リザバ層より前に必要な情報を抽出し，抽象化するための層が必要になる．しかし，強化学習アルゴリズムにより生成した教師信号に基いた誤差信号をリザバより下層に伝播して学習を行うことが困難であった．今後下層部分の学習を行う手法を検討する必要がある．

謝辞

本研究は JSPS 科研費（奨励研究）JP18H00543 の助成を受けた。

参考文献

[1] Y.LeCun, Y.Bengio, G.Hinton : Deep learning. Nature 521, 436-444 (2015)

[2] V.Mnih et al. : Playing Atari with deep reinforcement learning. arXiv preprint arXiv : 1312.5602. (2013)

[3] H.Jaeger. : The “echo state” approach to analysing and training recurrent neural networks - with an erratum note. Bonn, Germany: German National Research Center for Information Technology GMD Technical Report 148.34 (2001): 13.

[4] T Matsuki, and K Shibata : Reinforcement Learning of a Memory Task Using an Echo State Network with Multi-layer Readout. Int’l Conf. on Robot Intelligence Technology and Applications. Springer Cham 17-26 (2017)