多層リードアウトエコーステートネットを用いた記憶タスクの強化学習

大分大学　○松木俊貴　柴田克成

Reinforcement Learning of a Memory Task using an Echo State Network with Multi-Layer Readout

Toshitaka Matsuki and Katsunari Shibata, Oita University

Abstract: Recently, an approach of training a Neural Network(NN) through Reinforcement Learning(RL) has been focused on and a Recurrent NN(RNN) is used in current studies. On the other hand, to solve the difficulties in learning of an RNN, the Reservoir Network(RN), which is a special RNN, has attracted much attention owing to its rich dynamic representations. Aiming to acquire more complex representations, an approach of using a Multi Layer Readout(MLR), which consists of a Multi Layer NN, has been studied. We expect that an RN with MLR can acquire necessary functions such as memory through RL. This paper shows that an RN with MLR can learn "memory task", which requires memory function and non-linear transformation of outputs, through RL with Back Propagation. The result suggests that insufficient computational ability of randomly connected RN needs MLR to learn through RL.

1　序論

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てきている。このことは、柔軟かつ並列な Neural Network(NN) が、学習により、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唆している。さらに近年、入力から出力までのすべての処理を NN により構成し、それを強化学習により学習する End-to-End RL の手法が注目されている [1]。我々のグループは、高次機能を持ったシステムの実現のために、上記のような手法をとる必要があると長年主張し研究を行ってきた [2]。最近では、DeepMind 社がこのような手法により Atari 社のゲームを学習させることに成功した [3]。この手法は、タスクに関する知識を与えることなく、行動の結果得られる報酬と罰のみを手がかりに、システムが合目的的で汎用的な内部表現やさまざまな機能を自律的に獲得することができる点で非常に優れている。

システムが時間の流れの中で適切な行動を学習するためには、時系列データの処理機能や必要となる内部ダイナミクスを獲得できなければならない。NN がそのような学習をする時、リカレント構造が必要になる。我々はリカレント NN(RNN) を強化学習により生成した教師信号に基づき、Back Propagation Through Time(BPTT) によって学習することで、記憶や予測といった機能が獲得できることを示してきた [4][5]。しかし、多段階に状態を遷移させていくような、複雑なダイナミクスを学習により獲得させることが困難であるという問題に直面してきた。

RNN を学習させるため BPTT が一般に広く用いられるが、RNN の学習には収束の遅さ、不安定さ、計算処理の複雑さ、といった様々な課題が存在する。そのような課題を回避する一つの手法として、Liquid State Machine[6] や Echo State Network(ESN)[7] のような Reservoir Network(RN) が用いられる。RN はランダムに決められ固定された重み値でスパースに相互結合したニューロンで構成されている"リザバ"と呼ばれる RNN をもつ。リザバは、入力やフィードバックされた出力を内部に取り込み、リッチな情報を保持したダイナミクスを形成する。RN の出力は Readout Unit(RU) によりリザバダイナミクスの線形和をとることにより生成され、望ましい値を出力するためにリザバから RU への重み値のみが学習される。そのため、RN は非常にシンプルな方法により簡単に時系列データの処理や、複雑な時系列パターンの生成を学習することができる。我々は、RN のようなカオティックなネットワークがもつリッチなダイナミクスから、必要なダイナミクスを抽出・再構成するようなアプローチが、前述の複雑なダイナミクスの学習による獲得という課題を解決する鍵になるのではないかと期待している。本研究では、レートモデルニューロンを持ち、制御 [8] やダイナミックパターン生成の学習 [9] など様々な研究で注目されている ESN を用いて検証を行う。

また、リザバが生成する信号と現在のセンサ信号の線形和だけでは表現しきれない出力を生成するため、通常用

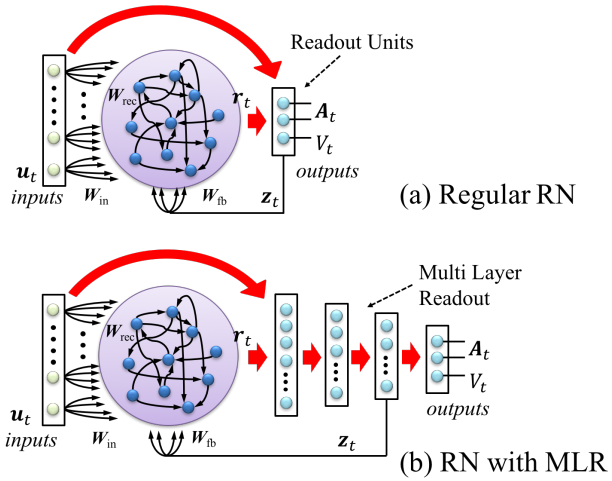


Fig 1: The network architectures of (a) a regular Reservoir Network(RN) and (b) an RN with Multi-Layer Readout(MLR).

いられる RU の代わりに、より表現力の高い Multi Layer Neural Network(MLNN) を用いた Multi Layer Readout(MLR) で出力を生成し Back Propagation(BP) 法で学習する研究が行われている [12]。Bush らは MLR をもつ ESN が部分観測環境下での Q-learning による Q 関数の近似が可能であることを示した [10]。また、Babinec らはこのような手法により時系列データ予測の精度が向上したことを示した [11]。このような研究は 10 年以上前に行われているが、我々は、このようなアーキテクチャが、End-to-End RL の重要性が高まりつつある中で、より一層複雑なダイナミクスおよび処理が求められるにつれ、今後さらに重要になってくると考えている

そこで、本論文では、過去の必要な情報を記憶し、それを活用して適切な行動を生成することを MLR をもつ ESN に学習させる。そして、そのような機能の学習が、過去に遡る処理が必要な BPTT ではなく、単に BP により学習することができることを示す。

2 研究方法

2.1 ネットワーク

Fig.1 に本研究で使用するネットワーク構造を示す。本研究では、RN は Fig.1(a) のような通常用いられる Readout Unit(RU) の代わりに、Fig.1(b) のような Multi Layer Neural Network (MLNN) を用いた Multi Lyaer Readout(MLR) で出力を生成する。タスク環境からの入力は、リザバと MLNN へと送られ、リザバ内部のニューロンの出力は MLNN へと送られる。リザバが持つ過去の沢山の情報を内部に保持する能力と、MLNN が持つ柔軟に沢山の情報から必要な情報を抽出し出力を作り出す能力を組み合わせる。このような構成により、信号を記憶し適切な出力を生成するための機能を、BPTT のように過去の情報まで遡ることのない BP によって学習できることが期待される。

リザバ内部のニューロン数 N_x は 1000 個で、それらは全て動的モデルであり、結合確率 $p = 0.1$ でスパースに結合している。時刻 t におけるリザバ内のニューロンの内部状態ベクトル $\boldsymbol{x}_t \in \mathbb{R}^{N_x}$ は次式により与えられる、

$$\boldsymbol{x}_t = \Big(1-a\Big)\boldsymbol{x}_{t-1} + a\Big(\lambda \boldsymbol{W}_{rec}\boldsymbol{r}_{t-1} + \boldsymbol{W}_{in}\boldsymbol{u}_{t-1} + \boldsymbol{W}_{fb}\boldsymbol{z}_{t-1}\Big) \tag{1}$$

ここで、$a = 0.1$ はリザバ内部のダイナミクスのスピードを決定する leaking rate である。$\boldsymbol{W}_{rec} \in \mathbb{R}^{N_x \times N_x}$ は、リザバニューロンの相互結合重み値行列であり、$\boldsymbol{r}_t$ はリザバ内部のニューロンの出力である。$\boldsymbol{W}_{rec}$ の値は平均 0、分散 $1/pN_x$ のガウス分布によってランダムに決定する。$\lambda = 1.2$ はリザバ内部の相互結合の重み値のスケールを決めるパラメータであり、この値が大きいほどリザバ内部のダイナミクスはよりカオティックになる。$\boldsymbol{W}_{in} \in \mathbb{R}^{N_x \times N_i}$ は、入力 $\boldsymbol{u}_t$ からリザバ内部のニューロンへの重み値行列であり、$\boldsymbol{W}_{fb} \in \mathbb{R}^{N_x \times N_f}$ は、MLR からのフィードバックベクトル $\boldsymbol{z}_t$ とリザバ内部のニューロンとの結合重み値行列である。ここで、N_i, N_{fb} は入力の数及び MLR からリザバへのフィードバックの信号数である。$\boldsymbol{W}_{in}$ と $\boldsymbol{W}_{fb}$ の値は -1 から 1 の一様乱数によって決定される。リザバ内の全てのニューロンの活性化関数は $tanh$ 関数である。

MLR は 4 層の NN で、最下層から順に $100, 40, 10, 3$ 個の静的ニューロンを持つ。MLR 内の全てのニューロンの活性化関数は $tanh$ 関数である。リザバ内の全てのニューロンの出力は、MLR の最下層ニューロンと全結合している。また、出力層一つ手前の中間層の $N_f = 10$ 個のニューロンの出力 $\boldsymbol{z}_t \in \mathbb{R}^{N_f}$ は、リザバ内の全てのニューロンへとフィードバックされる。環境から得られる $N_i = 7$ つの入力が存在し、それぞれがリザバ内のニューロン及び MLR の最下層ニューロンと全結合している。

本研究では、Actor-Critic により強化学習を行う。ネットワークは Critic 信号 V_t と Actor 信号 $\boldsymbol{A}_t$ を出力する。Actor の出力 $\boldsymbol{A}_t$ に, 探索成分ベクトル $\boldsymbol{rnd}_t$ を加えたものが、エージェントの時刻 t における動作信号となる。$\boldsymbol{rnd}_t$ の全ての値は -1 から 1 の一様乱数である。

2.2 学習方法

本研究では、ネットワークは Fig.1 の赤い矢印で示された部分の重み値に限り、強化学習に基づいて生成された教師信号と BP 法によって学習する。前の時刻 $t-1$ の

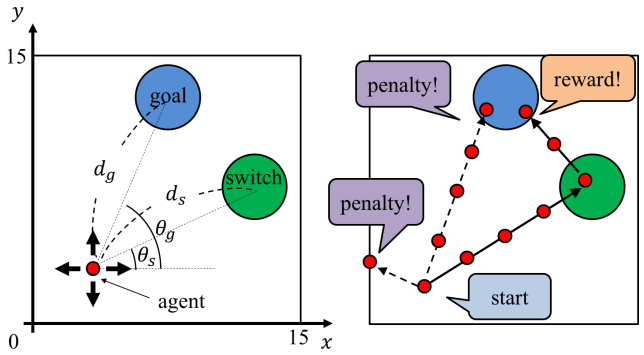


Fig 2: Outline of the memory task. An agent has to go to and enter the switch area at first, and then go to the goal area.

Critic の教師信号は次の式のように与えられる。

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

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。

$$\hat{r}_{t-1} = r_t + \gamma V_t - V_{t-1} \tag{3}$$

r_t は、時刻 t における環境からの報酬である。$\gamma = 0.99$ は割引率である。Actor の教師信号は次式により得られる。

$$\boldsymbol{A}_{t-1}^{teacher} = \boldsymbol{A}_{t-1} + \hat{r}_{t-1}\boldsymbol{rnd}_{t-1} \tag{4}$$

リザバが有する重み値 $\boldsymbol{W}_{rec}, \boldsymbol{W}_{in}, \boldsymbol{W}_{fb}$ は固定されており、学習をおこなわない。

2.3 タスク環境

MLR を持つ RN の能力を確かめるため、記憶を必要とするスイッチタスクを用いて検証を行う。通常の RN と比較することにより、MLNN の並列かつ柔軟な処理能力が記憶タスクにおいて有用であるかを確かめる。

タスクの概要を Fig.2 に示す。エージェントは縦横 15.0 の大きさを持つ二次元の平面上に置かれ、毎ステップ x 軸方向と y 軸方向への連続値の移動距離を決定する Actor 出力にしたがって移動する。エージェントはスイッチエリアに入った後、ゴールエリアに向かうという行動を学習する。エージェントは大きさを持っておらず、それぞれのエリアの半径は 1.5 である

エージェントは環境から 7 つの情報からなる入力ベクトル $\boldsymbol{u}_t = [d'_g, sin\theta_g, cos\theta_g, d'_s, sin\theta_s, cos\theta_s, signal]$ を受け取る。ここで、d'_g, d'_s はゴールおよびスイッチの中心までの距離 $d_g, d_s([0, 15\sqrt{2}])$ を $[-1, 1]$ の区間へと線形変換した値であり、θ_g, θ_s はエージェントから見た時の x 軸方向とゴールおよびスイッチの中心へと引いた直線とがなす角度である。また、$signal$ はエージェントがスイッチ上にいる間のみ得られる信号であり、次式に従う

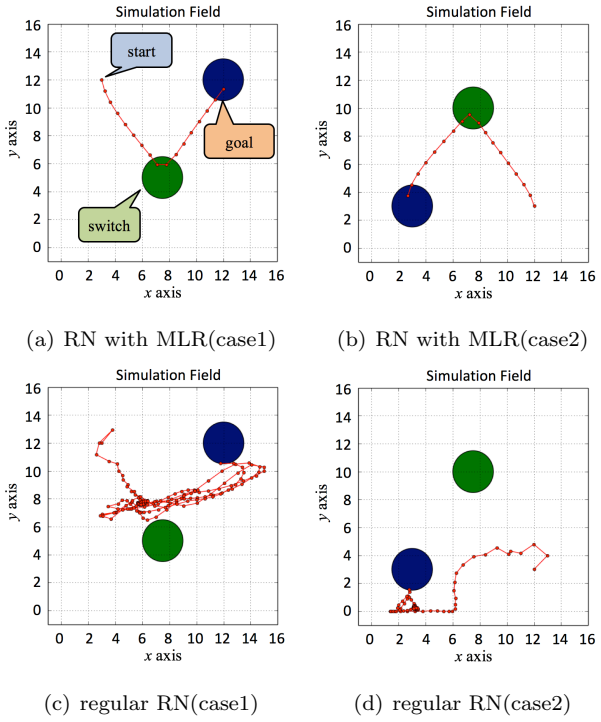
$$signal = \begin{cases} 0 & d_s > R_s \\ 10 & d_s \leq R_s \end{cases} \tag{5}$$

ここで、R_s はスイッチの半径である。1 試行ごとにエージェントの初期位置、ゴールエリア位置、スイッチエリア位置がフィールド内にランダムに設定される。この時、エージェント初期位置と各エリア範囲はそれぞれお互いに重なることはない。エージェントは、フィールド端の壁に接触すると $r_t = -0.1$ の罰を、スイッチエリアを経由せずにゴールエリアに入ると $r_t = -0.5$ の罰を、スイッチエリアを経由してゴールエリアに到達すると $r_t = 0.8$ の報酬をそれぞれ受け取る。1 試行はエージェントが 200 ステップ行動を行うか、ゴールエリアに入り報酬か罰を受け取ったときに終了する。エージェントが 50,000 試行学習を行った後、学習を停止し、テストを行う。

3 結果

エージェントに対し 50,000 試行の学習を行わせた後、エージェント初期位置、スイッチ位置、ゴール位置を、スイッチを押す前後でエージェントが出力すべき Actor の値が逆になるような 2 パターンに設定し、テストを行った。この時の、エージェントの軌道を Fig.3 に示す。比較のため、MLR を持つ RN の学習結果を Fig.3(a)(b) に、RU を持つ通常の RN の学習結果を Fig.3(c)(d) に示した。Fig.3(a)(b) に示すように、MLR を持つ RN のエージェントは始めにスイッチへと向かい、スイッチを押した後ゴールへ向かっている。このことから、このネットワークは、BPTT による学習のように過去に遡ることなく、単純な BP のみによって記憶の機能と適切な動作信号生成の機能を獲得し、記憶タスクの強化学習に成功していることが分かる。一方 Fig.3(c)(d) より、通常の RN のエージェントは、学習がうまくいかず、スイッチとゴールの間をさまよったり、壁の方向へぶつかり続けたりした後、スイッチを押さずにゴールしている。

適切な出力を生成するためには、現在のセンサ信号とスイッチエリアに入ったかどうかを記憶した情報との非線形な統合が必要になる。通常の RN がこのタスクを学習することができなかったことを考慮すれば、タスクに必要となるスイッチ前後での出力の非線形な転換は、リザバのみでは生成できず、MLR の学習によって構成されたと考えられる。このことから、RU よりも表現力の高い MLR がリザバの出力とセンサ信号を非線形に統合し、記憶に基づいて Actor 出力を切り替えるために必要となると考えられる。



(a) RN with MLR(case1) (b) RN with MLR(case2)

(c) regular RN(case1) (d) regular RN(case2)

Fig 3: Comparison of agent trajectory for two cases between RN with MLR and regular RN.

4 結論

出力を RU の代わりに MLNN で生成する MLR をもった RN が記憶タスクを BPTT ではなく BP により強化学習できることを示した。スイッチを押したことを記憶し、スイッチを押す前後で出力を非線形に転換する働きは、MLR が入力とリザバダイナミクスを統合し、適切な Actor 出力の生成をおこなうことにより実現していると考えられる。今後の課題として、更に複雑なタスクへの適用を試みること、ネットワーク内の記憶がどれだけの期間保持されうるかの分析、MLR の中間層のフィードバックがリザバダイナミクスにどのような影響を与えているかの分析、必要な情報を抽出するリザバの入力重み値の学習方法の検討、リザバが持つカオティックな内部ダイナミクスを利用した探索による我々の提案する新しい強化学習手法 [13][14] が適応可能かどうかの検証などが挙げられる。

謝辞

本研究は JSPS 科研費 (15K00360) の補助を受けた。

参考文献

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

[2] 柴田克成: 深層学習が示唆する end-to-end 強化学習に基づく機能創発アプローチの重要性と思考の創発に向けたカオスニューラルネットを用いた新しい強化学習 : 「認知科学」(Vol. 24, No.1, pp. 96-117) (2017)

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

[4] K.Shibata and H.Utsunomiya : Discovery of Pattern Meaning from Delayed Rewards by Reinforcement Learning with a Recurrent Neural Network, Proc. of IJCNN., pp. 1445-1452(2011)

[5] K.Shibata and K.Goto : Emergence of Flexible Prediction-Based Discrete Decision Making and Continuous Motion Generation through Actor-Q-Learning, Proc. of ICDL-Epirob. 2013, ID 15 (2013)

[6] 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.

[7] W.Maass, T.Natschlger, and H.Markram. : Real-time computing without stable states: A new framework for neural computation based on perturbations. Neural computation 14.11, 2531-2560.(2002)

[8] M.Salmen, and P.G.Ploger. : Echo state networks used for motor control. Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on. IEEE (2005)

[9] D.Sussillo, L.F.Abbott : Generating coherent patterns of activity from chaotic neural networks. Neuron Article, Vol.63, No.4, pp.544-557(2009)

[10] K.Bush, and C.Anderson. : Modeling reward functions for incomplete state representations via echo state networks. : Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on. Vol. 5. IEEE (2005)

[11] Š.Babinec, and J.Pospchal. : Merging echo state and feedforward neural networks for time series forecasting. : Artificial Neural NetworksICANN 2006, pp.367-375. (2006)

[12] M.Lukusevičius and H.Jaeger. : Reservoir computing approaches to recurrent neural network training. : Computer Science Review 3.3, pp.127-149. (2009)

[13] Y.Goto and K.Shibata : Emergence of Higher Exploration in Reinforcement Learning Using a Chaotic Neural Network, Proc. of Int'l Conf. on Neural Information Processing (ICONIP)2016, LNCS 9947, pp. 40-48 (2016)

[14] T.Matsuki and K.Shibata : Reward-Based Learning of a Memory-Required Task Based on the Internal Dynamics of a Chaotic Neural Network, Proc. of Int'l Conf. on Neural Information Processing (ICONIP)2016, LNCS 9947, pp. 376-383 (2016)