多層リードアウトを持つリザバを用いた強化学習におけるネットワーク構造の検討

大分大学　○吉岡晴海　松木俊貴　柴田克成

Examination of network structure in Reinforcement Learning using a Reservoir Network with Multi-Layer Readout

Harumi Yoshioka, Toshitaka Matsuki and Katsunari Shibata, Oita University

Abstract: Our group has studied to use a reservoir network (RN), which has rich chaotic dynamics, in reinforcement learning (RL), and showed that the learning performance in memory-required task was largely improved by introducing a multi-layer readout (MLR) instead of a single layer readout. In this paper, We compare the learning performance of switch task that requires memory by changing the number of layers of MLR, the position of feedback, the output to be feedback, and recurrent connection weight value scale λ. The results show that MLR can learn with two layers at least, feedback is needed to learn through RL and feedback from higher layers is effective. Learning through RL requires feedback, and the recurrent connection weight value scale in the reservoir is optimal around $\lambda = 1.5$.

1　序論

多くの分野で Deep Learning によるアプローチの有用性が注目されている。われわれの研究室では、以前より自律的な高次機能の創発により、脳のようなシステムの実現を目指してきた。そのために、入力から出力までをニューラルネットワーク (NN) で構成し、その学習に強化学習を用いる End-to-End 強化学習を研究してきた [1]。この手法では、探索で得られる報酬と罰のみで、システムが様々な機能を自律的に獲得することができる。近年では DeepMind が、同様のアプローチで TV ゲームやボードゲームなどで驚くべき結果を示している [2][3]。

音声認識や自然言語処理などの時系列データを扱うケースでは、リカレント NN(RNN) が用いられる [4]。また、われわれは End-to-End 強化学習に RNN を用いることで、時間的処理を必要とする記憶や予測といった機能の創発を確認した [5][6]。高次機能の典型である「思考」には多段階な内部の状態遷移が必要である。しかし、RNN では限定された状況で数回の状態遷移を学習することが限界であった [7]。

RNN の学習法には BPTT(Back Propagation Through Time) が広く用いられているが、BPTT を用いた学習には、計算量が多い、学習が不安定であるといった問題点がある。その問題の解決策として、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、疎に相互結合をしたニューロンで構成されたリザバと呼ばれる RNN を持つ。リザバは内部ダイナミクスの中に入力情報を長期間保持できる。そのため、出力層がリザバから必要な情報を取り出すために、出力層の重み値のみを学習するだけで通常の RNN より簡単に時系列データを扱うことができる [8]。

われわれは、RN が内部に持つカオスダイナミクスを利用すれば、通常の RNN よりも多段階の状態遷移をするための複雑なダイナミクスの形成に有利なのではないかと考えた。そこでまず、RN に強化学習で記憶が必要なスイッチタスクを行わせたが、単層の Readout Unit では学習ができなかった。一方、出力層を階層型 NN に置き換え、Multi-Layer Readout(MLR) とした RN with MLR では、時間を遡らない BP(Back Propagation) を適用するだけで学習ができた [9]。そこで、本論文では、この MLR のフィードバックを中心に、ネットワーク構造やパラメータが学習にどう影響を与えるか調査し、何が学習性能に大きく影響するかを探ってく。

2　研究方法

2.1　ネットワーク

Fig.1 に本研究で使用する Reservoir Network(RN) with Multi-Layer Readout(MLR) のネットワーク構造を示す。入力はリザバと MLR に入り、リザバ内ニューロンの出力は MLR へと送られる。MLR は外部からの入力とリザバの出力から最終的な出力を計算する。リザバ内は $N = 200$ 個の動的モデルニューロンで構成され、結合確率 $p = 0.1$ で疎結合している。ある時刻 t でのリザバ内ニューロンの内部状態ベクトル $\boldsymbol{x_t}$ は次式により与えられる、

$$\begin{aligned} \boldsymbol{x}_t &= \left(1-a\right)\boldsymbol{x}_{t-1} \\ &+ a\left(\lambda \boldsymbol{W}^{\mathrm{rec}}\boldsymbol{r}_{t-1} + \boldsymbol{W}^{\mathrm{in}}\boldsymbol{u}_t + \boldsymbol{W}^{\mathrm{fb}}\boldsymbol{z}_{t-1}\right) \end{aligned} \quad (1)$$

ここで、$a = 0.5$ 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。$\boldsymbol{W}^{\mathrm{rec}}$ 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

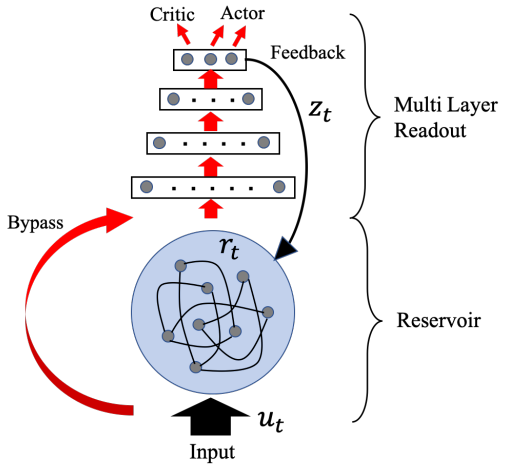

Fig. 1: The network architectures of Reservoir Network with Multi-Layer Readout(RN with MLR).

ューロンの出力である。$\boldsymbol{W}^{\mathrm{rec}}$ の値は平均 0、分散 $1/pN$ のガウス分布によりランダムに決定する。$\lambda = 1.2$ はリザバニューロンの相互結合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。$\boldsymbol{W}^{\mathrm{in}}$ は、リザバのニューロンへの入力重み値行列であり、$\boldsymbol{W}^{\mathrm{fb}}$ 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。$\boldsymbol{W}^{\mathrm{in}}$ と $\boldsymbol{W}^{\mathrm{fb}}$ の値は -1 から 1 の一様乱数によってランダムに決定する。

リザバ内の全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おり、出力層の出力 $\boldsymbol{z}_t$ は、リザバ内の全ニューロンへとフィードバックされる。環境から得られる入力 $\boldsymbol{u}_t$ は、リザバ内ニューロンと MLR の最下層ニューロンに全結合している。MLR は静的ニューロンで構成され、前の時間の情報を保持せず、時間毎に内部状態が入れ替わる。また、リザバ、MLR 共に全てのニューロンの活性化関数は tanh 関数である。

ここでは強化学習でよく用いられる Actor-Critic と呼ばれる構成を使用する。ネットワークは状態評価値 Critic の V_t と動作出力ベクトル Actor の $\boldsymbol{A}_t$ を出力する。エージェントの動作信号は動作出力 $\boldsymbol{A}_t$ に探索成分ベクトル $\boldsymbol{rnd}_t$ を加えたものとする。$\boldsymbol{rnd}_t$ は -1 から 1 の一様乱数でランダムに決まる。

2.2 学習方法

本研究では、強化学習に基づいて生成された教師信号を用いて、BP 法によって Fig.1 に示すネットワークの赤い矢印部分の重み値のみを、毎ステップ行動するたびに更新する。その際の学習係数はいずれも 0.01 である。

状態評価値 Critic の教師信号 T_V は次の式で決まる。

$$\begin{aligned} T_{V_{t-1}} &= V_{t-1} + \hat{r}_{t-1} \\ &= r_t + \gamma \cdot V_t \end{aligned} \tag{2}$$

$\hat{r}_{t-1}$ は時刻 $t-1$ での TD 誤差で、次の式で求められる。

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

r_t は時刻 t で獲得する報酬で、γ=0.96 は割引率である。動作出力 Actor の教師信号は次の式で求める。

$$\boldsymbol{T}_{A_{t-1}} = \boldsymbol{A}_{t-1} + \boldsymbol{rnd}_{t-1} \cdot \hat{r}_{t-1} \tag{4}$$

2.3 タスク設定

記憶が必要な、スイッチタスクを用いて学習の性能を比較する。タスクの設定を Fig.2 に示す。エージェントは初め、大きさが縦横 15.0 の二次元の平面上にランダムに置かれ、毎ステップ x 軸方向と y 軸方向への動作出力 Actor に探索成分の乱数を加えた分だけ移動する。エージェントはスイッチを押した後ゴールに入れば報酬 $r_t = 0.9$ を得ることができ、壁にぶつかると罰 $r_t = -0.1$ が、スイッチを踏まずにゴールに到達すると罰 $r_t = -0.9$ が与えられ 1 試行が終了する。

エージェントに大きさはなく、スイッチとゴールの半径はそれぞれ $R_s = R_g = 1.5$ であり、エージェントは環境から入力ベクトル $\boldsymbol{u}_t$=$[d_g, \sin\theta_g, \cos\theta_g, d_s, \sin\theta_s, \cos\theta_s, S]$ 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各距離とフィールドの対角距離を用いて -1 から 1 の間に正規化したものである。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、S はエージェントがスイッチ上にいる間のみ入力される信号であり、次式に従う。

$$S = \begin{cases} 0 & (d_s > R_s) \\ 1 & (d_s < R_s) \end{cases} \tag{5}$$

毎試行エージェントの初期位置と重らないよう、ゴール位置、スイッチ位置をフィールド内にランダムに設定する。この時、1 試行はエージェントが上限である 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。

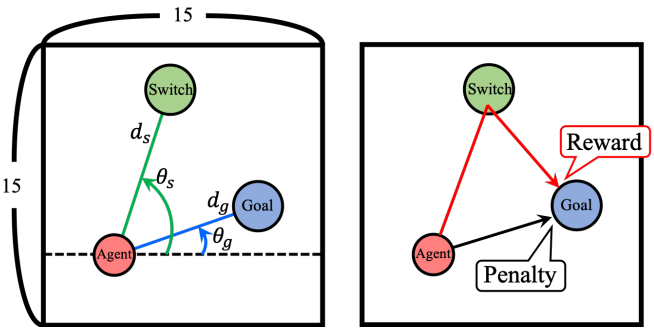

Fig. 2: Outline of the memory task.

3　結果

エージェントに 20,000 試行の学習を行わせ、その途中 2,500 試行毎に 30 パターンのフィールドでテストを行った。これを乱数系列 10 パターンで行い、スイッチを押してゴールにたどり着くのにかかったステップ数を記録する。この時、エージェントがスイッチを押さずにゴールに着いた場合は上限の 200 ステップが記録される。そして、テストを行う度にテストフィールドのパターン数と乱数系列でステップ数を平均し、学習性能の比較を行なった。

3.1　Multi-Layer Readout の層数の影響

Readout 部の層数が学習に与える影響を見るため、MLR の層数を 1 層から 4 層の間で変更した 5 つのネットワーク (1)〜(5) で学習性能の比較を行なった。それぞれの場合の各層のニューロン数は Table.1 に示した。また、結果のグラフを Fig.3 に示す。

Table 1: 5 networks used for comparison

	層数	ニューロン数
ネットワーク (1)	1 層	3
ネットワーク (2)	2 層	100-3
ネットワーク (3)	3 層	100-40-3
ネットワーク (4)		100-10-3
ネットワーク (5)	4 層	100-40-10-3

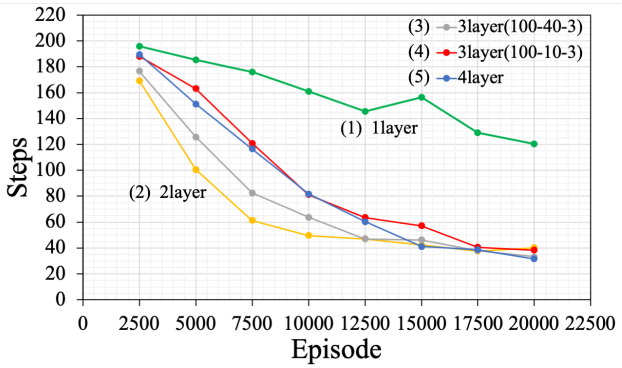

Fig. 3: Comparison of learning curves with different number of layers of MLR.

Fig.3 から、出力層が 1 層での学習はできておらず、2 層以上の時学習に成功していることがわかる。また、層数が増えるにつれ学習が遅くなっている。これは、層が多くなると誤差信号が伝わり難く、下層の学習が遅れるためだと考えられるが、より複雑な処理を必要とするタスクでは、多層の場合の方が有利になる可能性がある。

3.2　フィードバックを行う位置の影響

次に、フィードバックする情報のレベルが学習に与える影響を調べるため、4 層の MLR でフィードバックの位置を第 1 層から出力層で変更した。この時、層ごとにフィードバック重み値を変え、最も良い結果で学習性能の比較をした。その結果を Fig.4 に示す。

出力層またはその一つ下の第 3 層からフィードバックを行なった場合が、より良く学習ができていることがわかる。このことから、フィードバックは上位の層から行なうのが良いと考えられる。

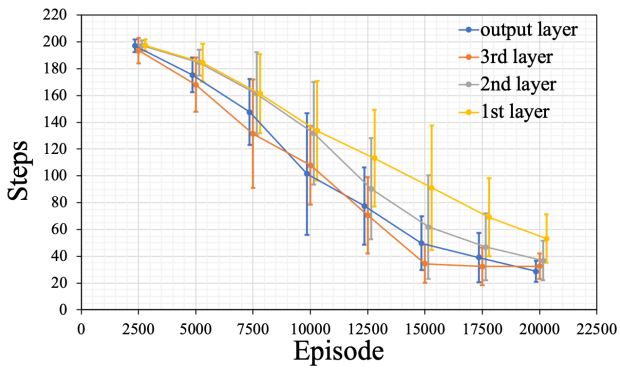

Fig. 4: Comparison of learning curves with different feedback positions of MLR.

3.3　フィードバックする出力の違いによる影響

出力は Actor と Critic の二つに分けられるが、そのうちどちらがフィードバックされる情報として学習に重要か確かめた。MLR 層数は 4 層とし、フィードバックを、出力層全体、Actor のみ、Critic のみ、フィードバックなしの 4 パターンで変更し、学習性能の比較を行なった。

Fig.5 に学習係数 0.01 とした時の結果を示す。出力層全体をフィードバックした結果と Actor のみをフィードバックした結果が同程度であり、Critic のみをフィードバックした結果とフィードバックを行わない場合の結果が同程度であった。この結果から、Critic はフィードバックする情報に必要ないようにも見える。しかし、学習係数を 0.05 に上げると、Fig.6 のような結果となった。このケースでは Fig.5 の場合とは逆に Critic のみをフィードバックした場合が Actor のみをフィードバックする場

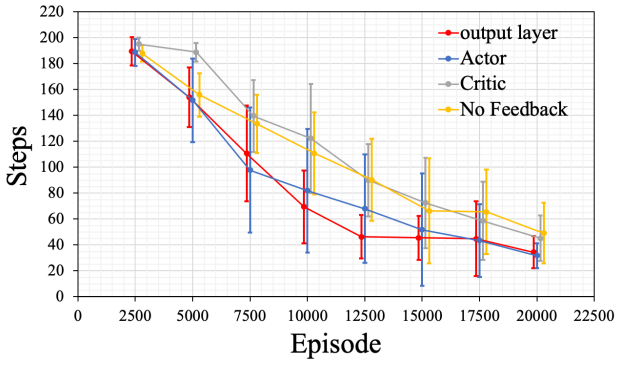

Fig. 5: Comparison of learning curve with different type of output feedback. (Learning rate:0.01)

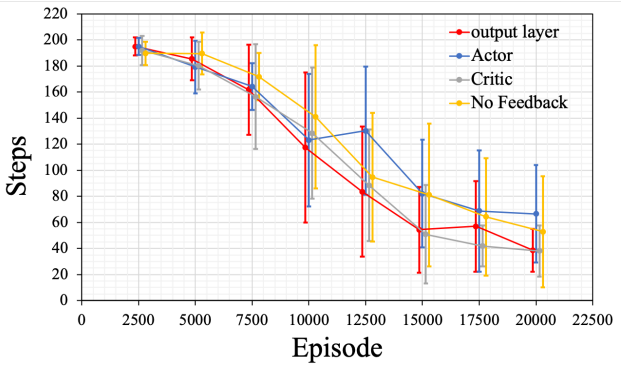

Fig. 6: Comparison of learning curve with different type of output feedback. (**Learningrate** : **0.05**)

合より学習性能が高くなった。

フィードバックがない場合の学習性能は一貫して悪いことから、フィードバックは RN with MLR の強化学習に必要であるということがわかる。しかし、Actor と Critic の個別のフィードバックについては、パラメータによって学習結果が大きく変わったことから、パラメータの最適化の問題であると考えられ、個々の影響を明確に比較することは困難であった。

3.4　リカレント結合重み値スケール λ の影響

最後に、リザバ内ダイナミクスのカオス性に関わる、リカレント結合重み値のスケール λ が学習に与える影響を調べた。Fig.7 に、$\lambda = 1.0$ から $\lambda = 2.5$ まで 0.5 間隔で λ を変更した学習結果を示す。

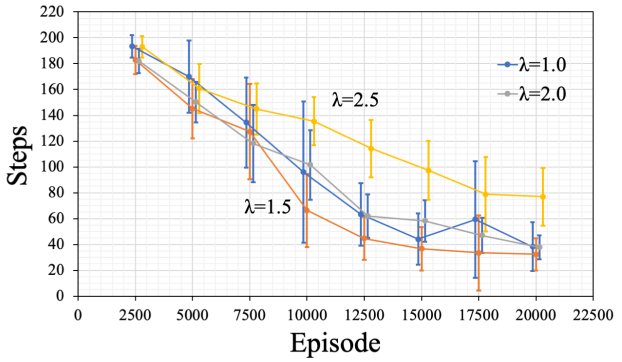

Fig. 7: Comparison of learning curves with different recurrent weight value scale.

この結果から、$\lambda = 1.5$ 付近で最もよく学習ができており、これまで用いた $\lambda = 1.2$ の場合より良かった。$\lambda = 1.5$ から値が変化すると、学習性能は低下傾向にある。このことから、大き過ぎず、小さ過ぎずない適切な値が $\lambda = 1.5$ 付近にあると言える。

4　結論

Reservoir Network(RN) の Readout Unit を階層型ニューラルネットに置き換えた RN with Multi-Layer Readout(MLR) を用いた強化学習において、MLR の層数とフィードバックを行う位置、フィードバックする情報、リカレント結合重み値スケール λ を変え、スイッチタスクでの学習性能がどのように変化するかを観察した。MLR の層数は少なくとも 2 層以上必要で、層数が少ないほどより早く学習ができた。フィードバックを行う位置は MLR の上位の層がよいことがわかった。フィードバックする情報については、動作と状態評価値どちらの情報も学習に有用である。λ については、これまで $\lambda = 1.2$ としてきたが、$\lambda = 1.5$ の方が良く、適切な値が $\lambda = 1.5$ あたりにあることがわかった。

謝辞

本研究は JSPS 科研費 (15K00360) の助成を受けたものである。ここに謝意を表する。

参考文献

[1] 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)

[2] V.Mnih, et al. : Human-level control through deep reinforcement learning, *Nature* Vol.518, pp.529-533 (2015)

[3] D.Silver, et al. : Mastering the game of Go with deep neural networks and tree search, *Nature* Vol.529, pp.484-489 (2016)

[4] A.Hannun, C.Case and J.Gasper et al. : DeepSpeech: Scaling up end-to-end speech recognition, *arXiv*, 1412.5567 (2014)

[5] 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)

[6] 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)

[7] Y.Sawatsubashi, M.F.Samsudin and K.Shibata : Emergence of Discrete and Abstract State Representation through Reinforcement Learning in a Continuous Input Task, *AISC Proc.of RiTA*2012, M1C-2.pdf, pp. 13-22 (2012)

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

[9] T.Matsuki and K.Shibata : Reinforcement Learning of a Memory Task using an Echo State Network with Multi-Layer Readout. Robot Intelligence Technology and Applications 5. (RiTA)2017. *AISC* Vol.751. Springer, pp. 17-26, (2017)