カオスニューラルネットの内部ダイナミクスを利用した記憶タスクの報酬に基づく学習

大分大学　○松木俊貴　柴田克成

Reward-Based Learning of a Memory-Required Task based on the Internal Dynamics of a Chaotic Neural Network

Toshitaka Matsuki and Katsunari Shibata, Oita University

Abstract: We have expected that dynamic higher functions such as "thinking" emerge through the growth from exploration in the framework of reinforcement learning(RL) using a chaotic Neural Network(NN). In this frame, the chaotic internal dynamics is used for exploration and that eliminates the necessity of giving external exploration noises. On the other hand, reservoir computing has shown its excellent ability in learning dynamic patterns. Hoerzer et al. showed that the learning can be done by giving rewards and exploration noises instead of explicit teacher signals. In this paper, aiming to introduce the learning ability into our new RL framework, it was shown that the memory-required task in the work of Hoerzer et al. could be learned without giving exploration noises by utilizing the chaotic internal dynamics while the exploration level was adjusted flexibly and autonomously.

1　序論

近年、大規模なニューラルネット (NN) に生のセンサ信号の超並列処理を学習させる DeepLearning によって、様々な分野で既存のシステムを凌駕する性能を獲得できることが示されている。このことは、並列処理を行う脳の驚異的な性能を “逐次的な意識” によって理解することの難しさと、脳のような超並列処理アルゴリズムを人の手で構築することの難しさを示唆している。我々のグループは、長年これらの難しさを指摘し、センサからモータまでの処理全体を NN で構築し、その中で必要な機能や意味のある内部表現が探索と報酬による強化学習によって発現するようなシステムの開発の必要性を主張してきた [1][2]。最近の研究では、ダイナミクスを扱うためにリカレント NN(RNN) を導入し、内部に “記憶” や “予測” の機能が創発することを簡単なタスクで確認した [3][4]。しかしながら、カオス性を持たない “silent” な RNN に学習を通じて多段階の状態遷移を形成することは困難であった [5]。

そこで我々は、複雑なダイナミクスは、“silent” な RNN に一から構築されるのではなく、カオス NN が持つリッチなカオスダイナミクスを元にして学習によって形成するべきと考えた。また、そのダイナミクスは強化学習の探索成分としても利用でき、外部からの探索ノイズの付加を不要にすることができる。そしてそれは、外界との因果関係を反映して合目的的なものへと変化し、最終的には複雑なダイナミクスが必要となる “思考” のような高次機能へとつながると考えている。この新しい強化学習手法により、いくつかの簡単なタスクの学習ができることを確認した [6][7]。

その一方で、近年、Echo State Network[8] や Liquid State Machine[9] のような、大規模な RNN が持つリッチなダイナミクスから必要なパターンを抽出することで、様々なタスクの学習を行うリザバーコンピューティング (RC) が注目されている。Sussillo らはリザバーネットワーク (RN) を FORCE Learning と呼ばれる新しい学習手法で学習させた [10]。この手法では、出力はネットワークへとフィードバックされ、学習中、常に出力が目標値に近似するように出力ユニットの重み値を修正する。これによって複雑なダイナミックパターンの生成を驚くほど簡単かつ急速に学習できることが確認された。

Hoerzer らは、明示的な教師信号を与える代わりに、探索ノイズと出力の誤差によって導かれるパフォーマンスの向上を示す報酬を与える Reward-Modulated Hebbian Learning によって、RN が学習できることを示した [11]。この研究では、様々なダイナミックパターンの生成や記憶タスクの学習が可能であることが示された。

以上のことから、我々は、“思考” のような高次機能を実現するため、我々が提案している前述の新しい強化学習手法に対して、それらのダイナミカルパターンの学習能力を導入することが必要であると考えた。本研究では、この試みの第一ステップとして、Hoerzer らの研究において報酬信号によって RN が学習した記憶タスクが、我々の新しい強化学習手法と同じように内部ダイナミクスを

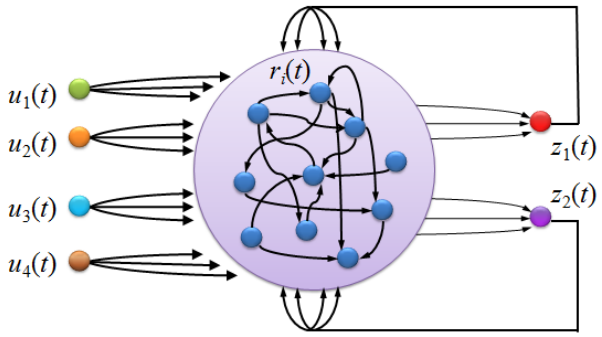


Fig 1: The network model. It has 4 inputs, u_1(green), u_2(orange), u_3(cyan), u_4(brown) and 2 outputs, z_1(red), z_2(purple). In the network, 1000 neurons(blue) are recurrently connected(connection probability $p = 0.1$).

探索に活用することで、外部からの探索ノイズなしに学習することができるかどうかを調査する。

2 研究方法

2.1 ネットワーク

本研究では、先行研究と同じ構造のネットワークを用いる（Fig.1）[10][11]。ネットワークのニューロン数 N は 1000 個で、それらはすべて動的モデルであり、結合確率 $p = 0.1$ で相互にスパース結合している。4 種類の外部入力が存在し、それらはすべてのニューロンに与えられる。2 つのリードアウトユニットと呼ばれる出力ユニットが存在し、それぞれがネットワーク内のすべてのニューロンと結合している。出力ユニットの出力 z_i はすべてのニューロンへとフィードバックされる。時刻 t における j 番目のニューロンの内部状態 x_j は次式で与えられる、

$$x_j(t) = \Big(1 - \frac{\Delta t}{\tau}\Big)x_j(t - \Delta t) + \frac{\Delta t}{\tau}\Big(\lambda \sum_{i=1}^{N} w_{ji}^{rec} r_i(t) + \sum_{i=1}^{I} w_{ji}^{in} u_i(t) + \sum_{i=1}^{O} w_{ji}^{fb} z_i(t)\Big) \tag{1}$$

ここで、シミュレーションステップ Δt は 1[ms]、時定数 τ は 10[ms] とする。また、λ はニューロンの相互結合の重み値のスケールを決定するパラメータであり、この値が大きいほどニューロンの活動はよりカオティックになる。本研究では λ を 1.8 に設定した。w_{ji}^{rec} は、i 番目のニューロンから j 番目のニューロンへの相互結合部の重み値であり、その値は平均 0、分散 $1/pN$ のガウス分布によってランダムに決定する。I は入力の数であり、w_{ji}^{in} は i 番目の入力から j 番目のニューロンへの重み値、u_i は i 番目の入力の値である。O はリードアウトユニットの数であり、w_{ji}^{fb} は i 番目のリードアウトユニットから j 番目のニューロンへの重み値である。w_{ji}^{in} および w_{ji}^{fb} はそれぞれ -1 から 1 の一様分布でランダムに決定する。各ニューロンの出力 $r_j(t)$ は、その内部状態 $x_j(t)$ から $tanh$ 関数を使って計算される。

$$r_j(t) = tanh\Big(x_j(t)\Big) \tag{2}$$

リードアウトユニットの出力 $z_j(t)$ は、各ニューロンの出力 $r_i(t)$ とそのニューロンとリードアウトユニットの間の重み値 w_{ji} によって以下のように計算される。

$$z_j(t) = \sum_{i=1}^{N} w_{ji} r_i(t) \tag{3}$$

w_{ji} の初期値は平均 0、分散 $1/N$ のガウス分布によってランダムに与えられる。

2.2 学習方法

本研究では、ニューロンとリードアウトユニットの結合の重み値 w_{ji} のみを学習する。探索ノイズを加えないということを除いて、基本的に Hoerzer らの学習手法に習って検証を行った [11]。ネットワークは、出力のパフォーマンスを表す $P(t)$ が、時定数 5[ms] での移動平均 $\overline{P}(t)$ と比べて向上したかどうかによって与えられる報酬と罰によって学習される。$P(t)$ は次のように定義される

$$P(t) = -\sum_{j=1}^{O}\Big(z_j(t) - f_j(t)\Big)^2 \tag{4}$$

ここで、$f_j(t)$ とは、j 番目のリードアウトユニットの目標出力である。ネットワークへの調整信号に関して、Hoerzer らの研究で用いられた Reward-Modulated Hebbian Learning を少し変更した。$P(t)$ と $\overline{P}(t)$ を用いて、調整信号 $M(t)$ を以下のように決定する。

$$M(t) = \begin{cases} 1 & P(t) > \overline{P}(t) \\ -1 & P(t) \le \overline{P}(t) \end{cases} \tag{5}$$

Hoerzer らの研究においては $M(t)$ は 0 か 1 の値をとっていたが、本研究では、$M(t)$ によって与えられる信号を報酬もしくは罰であると考え、1 か-1 と定義した。w_{ji} は $M(t)$ を使って次のように修正される。

$$\Delta w_{ji} = \eta\Big(z_j(t) - \overline{z}_j(t)\Big)M(t)r_i(t) \tag{6}$$

ここで η は学習係数であり、本研究では 0.0005 とした。$\overline{z}_j(t)$ は時定数 5ms での出力 $z_j(t)$ の移動平均である。

2.3 タスク

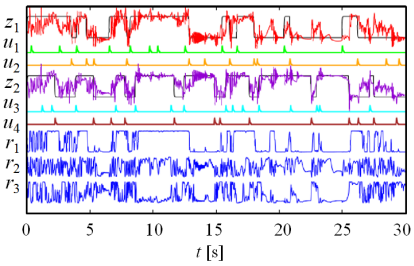
ネットワークは記憶を必要とするタスクの学習をおこなった [11]。ネットワークは 4 種類の入力と 2 つの出力を持つ。それぞれの入力は平均すると 2 秒に一度程度の間隔になる頻度でランダムに与えられる。その入力は 50[ms] かけて一次遅れで 1 まで上昇し、その後時定数 50[ms] で 0 まで減衰していくパルス入力である。各入力はそれぞれ異なった意味を持ち、u_1 と u_2 は出力 z_1 の ON 信号と OFF 信号であり、u_2 と u_3 は出力 z_2 の ON 信号と OFF 信号である。ON 信号は出力を 1 まで上げ、OFF 信号は出力を-1 まで下げる。出力が上がっている間に ON 信号が入力されるか、出力が下がっている間に OFF 信号が入力されるなどしても出力は変化しない。

3 結果

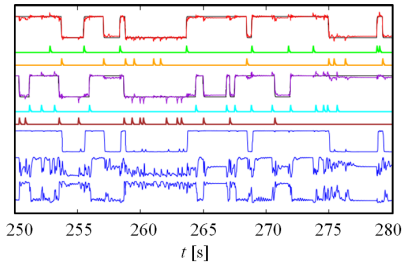
Fig.2 は 2 つのリードアウトユニットの出力 $z_{1,2}$、4 つのネットワークへの入力信号 $u_{1\sim4}$、3 つのニューロンの出力 $r_{1\sim3}$ の様子を示している。Fig.2(a) は学習開始から 30 秒間の様子を示している。開始直後は出力は目標値に全く追従していないが、外部からの乱数ノイズを加えていないにもかかわらず、ノイズのような激しい変動がネットワークの内部ダイナミクスから生まれ出力に現れていることが分かる。そしてその内部ダイナミクスが探索の役割を果たし、少し遅れて目標値へと追従を始めている。

Fig.2(b) は、ネットワークが 250 秒間学習を続けた後に学習を止め、30 秒間テストした様子を示している。出力はほとんど遅れることなく目標値に追従しており、このタスクが探索ノイズなしで学習可能であることが分かる。また、報酬と罰 ($M(t)$) によって学習が進行するに連れて、内部ダイナミクスによる出力の変動が徐々に低減していった。ネットワークは探索状態から安定状態へと遷移していったと考えられ、内部ダイナミクスによる探索成分は自律的に減少した。

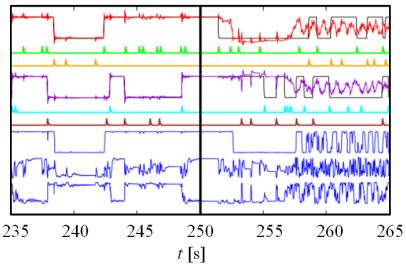
このネットワークが再び未知の環境に出会った時、探索レベルを自律的に調整することができるかどうかを観察するため、学習途中で急にタスクのルールを変更してみた。Fig.2(c) は、250 秒間ネットワークを学習させた後、u_1 と u_2 の働きの入れ替え及び u_3 と u_4 の働きの入れ替えをおこなった時のネットワークの様子を示している。ルールの変更後、外部からの指示を全く与えていないにも関わらず、再びカオティックな活動が現れ探索を再開していることが分かる。ルール変更前の 50 秒間と変更後に 300 秒間学習をさせた間の様子を、時間のスケールを圧縮して Fig.2(d) に示し、その後 30 秒間テストした様子を Fig.2(e) に示した。これらの結果は、たとえ学習の途中に突然環境が変化したとしても、ネットワーク



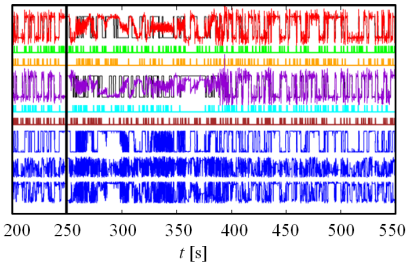
(a) First 30 seconds(training)



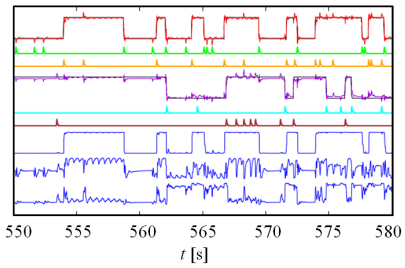
(b) 30 seconds (testing) after 250 seconds learning



(c) 30 seconds (training) around the rule change at 250 seconds



(d) 350 seconds (training) around the rule change at 250 seconds in a compressed time scale



(e) 30 seconds (testing) after 300 seconds from the rule change

Fig 2: Network activities. z_1, z_2 (red, purple). f_1, f_2 (black). u_1,u_2,u_3,u_4(green, orange, cyan, brown). The 3 sample of r_i(blue).

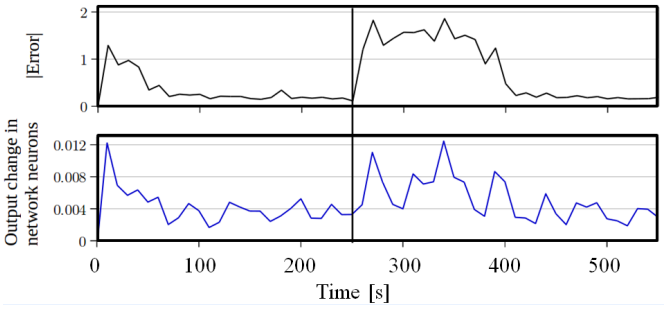


Fig 3: The mean absolute error(upper) and the mean absolute output change of network neurons(lower) during learning. The vertical line is the timing of rule change.

が探索を再開し、学習を成功させることができたことを示している。

学習の間、ネットワークニューロンのカオティックな活動がどのように変わっていくのかを示すため、出力の誤差とニューロンの出力の変化量を Fig.3 のように記録した。これらのグラフは、各ステップの出力誤差の絶対値平均と全ニューロンの出力の絶対値平均をそれぞれ求め、さらに 10000 ステップ毎にそれぞれ時間平均して記録したものである。Fig.3 のグラフから、始めのルールにしたがって学習を開始してから 250 秒の間出力誤差は減少していき、ネットワークの活動も徐々に低減していっていることが分かる。また、250 秒経過した時点でルールを変更した直後に出力誤差が増加し、少し遅れてネットワークの活動も活発になっていることが分かる。このことは、このネットワークが未知の状況に遭遇した時に、探索を行うために自律的に内部ダイナミクスをカオティックな状態に変えることが可能であることを示している。

4 結論

RN が報酬に基づいて記憶タスクを学習させる時、外部から加える探索ノイズの代わりに、ネットワークのカオティックな内部ダイナミクスが探索の役割を果たすことができることを確かめた。学習の進行と共に、内部ダイナミクスから発するノイズのような変動が低減していき、ネットワークの活動が探索状態から安定状態へと自律的に遷移していった。また、タスクのルール設定を学習の途中で変更した時、ネットワークがそれに呼応して探索と学習を再開することも確認した。これらの結果は、探索ノイズを用いない強化学習による RN の学習の可能性を示唆している。

謝辞

本研究は JSPS 科研費 (15K00360) の補助を受けた。

参考文献

[1] K.Shibata and Y.Okabe : Reinforcement Learning When Visual Signals are Directly Given as Inputs, Proc. of ICNN '97, Vol. 3, pp.1716-1720(1997)

[2] 柴田克成 : 強化学習とニューラルネットによる知能創発, 計測と制御, Vol. 48, No. 1, pp. 106–111, 2009

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

[4] 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)

[5] Y.Sawatsubashi, et al. : Emergence of discrete and Abstract State Representation in Continuous Input Task, Robot Intelligence Technology and Applications 2012, pp.13-22(2012)
沢津橋由人, 柴田克成 : リカレントネットを用いた強化学習における 離散的かつ抽象的な状態表現の創発, 計測自動制御学会 システム・情報部門学術講演会 2012 講演論文集, 3B1-3.pdf, pp. 402-407, 2012.11

[6] K.Shibata and Y.Sakashita : Reinforcement Learning with Internal-Dynamics-based Exploration Using a Chaotic Neural Network, Proc. of Int'l Joint Conf. on Neural Networks (IJCNN)2015, 2015.7
柴田克成, 坂下悠太: カオスニューラルネットを用いた内部ダイナミクス由来の探索に基づく強化学習, 電子情報通信学会技術報告, NC2014-117, pp.277-282, 2015

[7] Y.Goto and K.Shibata : Emergence of Higher Exploration in Reinforcement Learning using a Chaotic Neural Network, Proc. of ICONIP2016 (2016)

[8] H.Jaeger : The "echo state" approach to analysing and training recurrent neural networks. GMD Report 148, pp.43(2001)

[9] W.Maass, T.Natschlger, H.Markram:Real-time computing without stable states:a new framework for neural computation based on perturbations. NEURAL COMPUTATIOM, Vol.14, No.11, pp.2531-2560(2002)

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

[11] G.M.Hoerzer, R.Legenstein and W.Maass : Emergence of complex computational structures from chaotic neural networks through Reward-Modulated Hebbian Learning. Cerebral Cortex, Vol.24 No.3, pp.677-690 (2014)