強化学習によるリカレントニューラルネットワーク内部での振動子創発の可能性

大分大学　○品矢裕介　柴田克成

Emergence Possibility of Oscillator in a Recurrent Neural Network through Reinforcement Learning

∗Yusuke Shinaya, Katsunari Shibata ,Oita University

Abstract– It can be expected in a robot or agent that necessary functions emerge in a neural network through learning of action by reinforcement learning. Especially in recent years, in a dynamic environment, emergence of functions utilizing dynamics in a recurrent neural network has been expected. In this paper, emergence of oscillator as one of the typical dynamics is forcused on, and it is examined whether or not an oscillator emerges through learning of rythmic motion. A recurrent neural network learned periodic motion for forward-rotation in a piston-driven wheel by reinforcement learning with a reward given around a desired rotational speed. The system learned to roatate the wheel almost as desired. It was also confirmed that an oscillator emerged in the recurrent neural network by observing the wheel rotation after cutting off the inputs. The rotational speed was larger as the desired speed during learning was larger.

Key Words: reinforcement learning, recurrent neural network, oscillator

1　まえがき

情報であふれる実世界に住むわれわれ人間は、無数のセンサ信号を入力とした並列処理装置である脳の中で、認識、記憶、予測、行動計画などのさまざまな機能が調和的に働くことで、その時々の状況に応じた柔軟な行動を実現することができる。しかし、人間がこのようなロボットを開発しようとしても並列で柔軟なプログラムを組むことは難しいため、どうしてもそれぞれの機能モジュールに分けて考えてしまい、その結果、柔軟性を失ってしまうことになる。

これに対し、本研究室では、機能モジュールを 1 つ 1 つ組み込むのではなく、システム全体を学習可能な並列処理装置であるニューラルネットで構成し、試行錯誤することで報酬を得て罰を避ける行動を自律的に学習する強化学習を適用することで、ニューラルネット内部に必要な機能が創発することに注目して研究を進めてきた [1]。従来はセンサ-モータ間の静的なマッピングを学習するケースが多かった。しかし、われわれもロボットも時間の中で生活しており、ニューラルネット内部にダイナミクスを形成し、過去の情報を記憶したり、対象のダイナミクスを認識したり、何かを思考したりと、時間の流れを積極的に活用していくことでロボットの知的レベルを飛躍させることが求められる。そのような中で、リカレントニューラルネット (RNN) で過去の情報を考慮した行動を強化学習させることで、RNN 内部に固定点収束のダイナミクスを形成し、過去の必要な情報を抽出して記憶する機能を獲得できることは確認してきた。しかし、この枠組みでより複雑なダイナミクスが獲得できるかどうかはわかっていない。

より複雑なダイナミクスを RNN を用いた強化学習によって生み出すことができるかどうかを確認するため、本研究では振動子に注目する。振動子は生体内に存在し、歩行などのリズミックな周期運動をする際、環境の変化や外乱に対してロバストな運動を実現していると考えられている。ロボット等において、振動子を用いた歩行を学習によって獲得させる際には、通常、運動周期に近い周期を持つ振動子を予め組み込み、振動子自体は変化させず、外部とのインターフェイス部分のみ学習させることが多い [2]。ここではこの生体の重要な能力の一部と考えられている振動子を、強化学習によって周期運動を学習するうちに必要な機能として RNN 内部に創発するかどうかを確認する。

2　リカレントニューラルネットワークを用いた強化学習

本研究では Fig.1 のように RNN と強化学習を組み合わせた学習システムを用いる。RNN は階層型ニューラルネットをベースとし、最下層の中間層にフィードバックループを持たせた。文献 [3] では、強化学習は使用していないものの、連続時間型の RNN を用いて複雑なダイナミクスを学習によって形成しているため、本研究でも、細かい時間的変化を考慮し、学習できる連続時間型の RNN を用いる。連続時間型のニューロンの出力計算は、時刻 t における第 l 層の j 番目のニューロンの内部状態の値を $u_{j,t}^{(l)}$, 第 $l-1$ 層の i 番目ニューロンの出力を $x_{i,t}^{(l-1)}$、第 $l-1$ 層の i 番目のニューロンから第 l 層の j 番目のニューロンへの結合重みを $w_{ji}^{(l)}$、第 $l-1$ 層のニューロン数を $N^{(l-1)}$ とすると次の式 (1) で表される。

$$\tau_j \frac{du_{j,t}^{(l)}}{dt} = -u_{j,t}^{(l)} + \sum_{i=1}^{N^{(l-1)}} w_{ji}^{(l)} x_{i,t}^{(l-1)} \qquad (1)$$

τ_j は時定数である。また、上記の式を差分近似して整理すると n step 目の内部状態は以下の式となる。

$$u_{j,n}^{(l)} = (1 - \frac{\Delta t}{\tau_j}) u_{j,n-1}^{(l)} + \frac{\Delta t}{\tau_j} \sum_{i=1}^{N^{(l-1)}} w_{ji}^{(l)} x_{i,n}^{(l-1)} \quad (2)$$

Δt は刻み時間である。ここで、最下層の中間層ニューロン出力を、次の時刻の入力の一部として与えることでフィードバックループを形成した。強化学習としては、行動の連続値出力が可能な Actor-Critic を用いる。RNN の出力に動作生成部の actor1、 actor2、状態評価部である critic がそれぞれ割り当てられる。学習時は TD(Temporal Difference) 学習に基づいた強化学習のアルゴリズムからそれぞれの出力の教師信号を生成し、BPTT(Back Propagation Through Time) 法を用いて RNN の結合重みを更新していく。

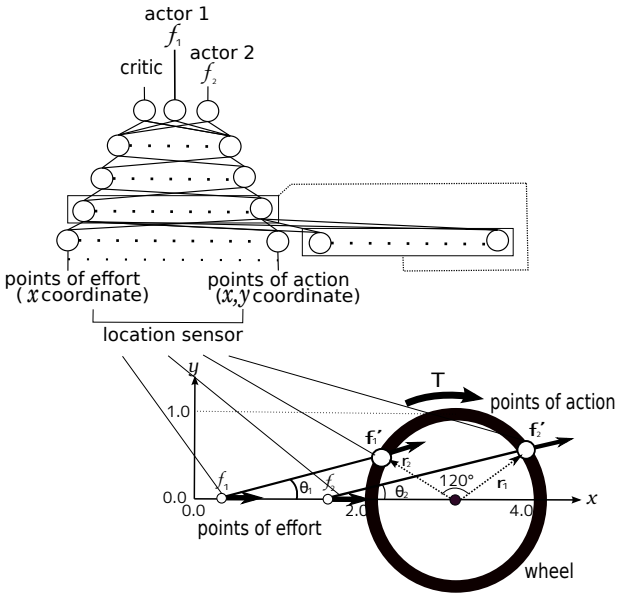


Fig. 1: Piston-driven wheel system and a RNN

時刻 t での critic $C(\mathbf{x}_t)$ に対する教師信号は、与えられた報酬 r_{t+1} および時刻 $t+1$ での環境の状態 $\mathbf{x}_{t+1}$ での critic $C(\mathbf{x}_{t+1})$ を用いて、

$$C_d(\mathbf{x}_t) = r_{t+1} + \gamma C(\mathbf{x}_{t+1}) \tag{3}$$

とする。ここで γ は割引率を表す。一方、actor $\mathbf{A}(\mathbf{x_t})$ の教師信号 $\mathbf{A}_d(\mathbf{x}_t)$ は、探索のために $\mathbf{A}(\mathbf{x}_t)$ に加えられる乱数である試行錯誤量 $\mathbf{rnd_t}$ と TD 誤差 $TDerror_t$ と actor の学習係数 α の積を用いて、

$$\mathbf{A}_d(\mathbf{x}_t) = \mathbf{A}(\mathbf{x}_t) + \alpha \times \mathbf{rnd_t} \times TDerror_t \tag{4}$$

$$TDerror_t = r_{t+1} + \gamma C(\mathbf{x}_{t+1}) - C(\mathbf{x}_t) \tag{5}$$

とし、critic $C(\mathbf{x}_t)$ の値がより大きくなるような動作を学習させる。以上の教師信号を用いて、RNN の結合重み値を BPTT 法に基づいて更新する。

3 タスク設定

Fig.1 に学習させるピストン駆動方式の車輪回転タスクを示す。車輪に 120 度の間隔を空けてロッドが 2 本付いており、そのロッドの両端に実際に力を加える力点 (x 軸上に拘束) と車輪との接続部の作用点がある。車輪の直径は 2.0m、ロッドの長さと力点の可動域も同様に 2.0m である。Fig.1 のように両ロッドの力点に x 軸方向の力 f_1,f_2 をそれぞれ加える。力 f_1 によって車輪の作用点にはロッドの角度 θ_1 に応じて、以下のような x 成分と y 成分を持つ力 $\mathbf{f}_1'$ がかかる。

$$f_{1,x}' = f_1 \cos^2(\theta_1), \quad f_{1,y}' = f_1 \cos(\theta_1)\sin(\theta_1) \tag{6}$$

$\mathbf{f}_2'$ も f_2 から同様にして求める。また、トルク $\mathbf{T}$ は $\boldsymbol{f}'$ とそれぞれの位置ベクトル $\boldsymbol{r}$ から以下の式で求められる。

$$\boldsymbol{T} = \boldsymbol{r}_1 \times \boldsymbol{f}_1' + \boldsymbol{r}_2 \times \boldsymbol{f}_2' \tag{7}$$

そして、車輪はトルク $\mathbf{T}$ によって以下の運動方程式に従って回転する。

$$I\frac{d^2\boldsymbol{\theta}}{dt^2} + D\frac{d\boldsymbol{\theta}}{dt} = \boldsymbol{T} \tag{8}$$

ここで、I は慣性モーメント、D は粘性摩擦係数、$\boldsymbol{\theta}$ は車輪の回転角度ベクトルである。この微分方程式をここでは差分で解くことで、次時刻の回転角加速度、角速度、角度を求める。

このタスクでは車輪を前進回転させるように RNN が学習する。RNN の入力には 2 つの力点の x 座標と 2 つの作用点の x 座標、y 座標の 4 点の位置情報を与える。この入力に対する強い非線形性を学習によって容易に実現できるように、これらの連続値信号の個々を Fig.2 のように 21 個の局所的に反応する信号で表現して RNN に入力する。

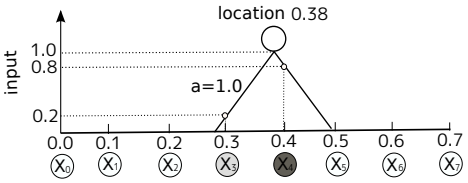


Fig. 2: Localized input signals

そして、この入力から RNN の出力として 2 つの actorf_1、f_2 と現在の評価値である critic の 3 出力を計算する。報酬には 1step 当たりの車輪の回転速度を利用する。学習当初は 100 試行の平均に 5 度足した値を目標回転速度として、最終目標回転速度に至るまで徐々に変化させる。報酬は回転速度が目標回転速度と一致した時にピーク値の 1.0、目標回転速度から 5.0 以上離れると 0 となるよう設定した。逆回転すると、その角度を 50 度で割った値を罰として与える。よって、RNN は報酬と罰によって目標回転速度を目指して、逆回転を避けるような学習を行う。学習終了の判定は、学習の進行を表すカウンターによって行う。このカウンターは学習始めは最終目標回転速度から ±(最終目標回転速度) の範囲なら +1、範囲外だと −1 毎ステップ変化させた。学習終了に近づくにつれてこの範囲を ±1 まで減少させ、カウンタの値が 10000 回になるとタスクを終了する。タスクとネットワークのパラメータを Tab.1 に示す。

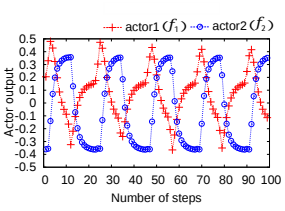
Table 1: Task Parameter setting

Wheel radius	$1.0m$
Inertia moment I	0.2 kgm^2
Kinematic viscosity coefficient D	0.01 Ns
Discount factor γ	0.9
Cofficient for actor learning α	0.9
Range of exploration rnd_x, rnd_y	
-1.0∼1.0 → -0.1∼0.1	
Number of neurons in each layer	
(Input)-120-50-30-20-3(Output)	
Value range of sigmoid function	-0.5∼0.5
Traced-back time in BPTT	30
Inital connection weight	
input-hidden1, hidden1-hidden2	-0.1∼0.1,-0.1∼0.1
hidden2-hidden3, hidden3-output	-0.1∼0.1,0.0
self-feedback, other-feedback	-0.1∼0.1
Learning rate	
For feedback connections	0.1
For other connections	0.1
Δt / τ	1.0/3.0

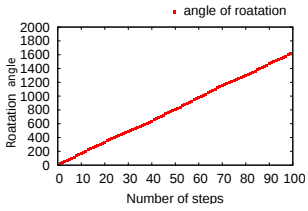
4 学習結果

最終的な目標回転速度が 10度/秒、15度/秒、20度/秒 の 3 つの場合のそれぞれについて、乱数系列を変化させて 5 パターン学習を行った。10度/秒 の時はすべて学習が終了したが、15度/秒､20度/秒 ではそれぞれ 1 パターンだけ学習できなかった。そして、学習した RNN を使って再びタスクのテストを行った。車輪が止まった状態から 500step 回転させた。入力を始めると Actor は周期的な力を出力し始め、車輪もすぐに回転し、学習した目標回転速度に近い速度で回転した。ただし、テスト開始時の RNN の初期の内部状態を調整しないと、うまく回転しない場合もあった。

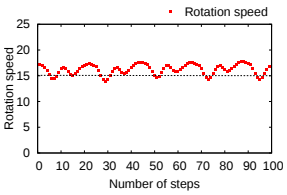
まず、目標回転速度が 15度/秒、20度/秒 で学習させた結果を 1 パターンずつ、それぞれ Fig.3、Fig.4 に示す。目標回転速度が 15度/秒 の時、30,400,972step、20度/秒 では 130,507,678step と長時間かけて学習が終了した。テストを行った時の 100 から 200step の間の Actor の出力、車輪の角度の変化、角速度および 2 点の力点の位置の変化をそれぞれ Fig.3 と 4 の (a) から (d) に示す。Fig.3、Fig.4 の (a) を見ると、Actor 出力の変化が Fig.3 では滑らかでなく、Fig.4 では滑らかとその形は大きく異なるものの、周期的に変化していることがわかる。Fig.3、Fig.4 の (b) を見ると、両者間で Actor 出力が大きく異なっていたものの、車輪の角度はいずれの場合も直線に近い変化をしている。Fig.3、Fig.4 の (c) を見ると、いずれも多少の変動はあるものの、概ね Fig.3 の場合が目標回転速度 15度/秒、Fig.4 の場合が 20度/秒 に近いところで推移していることがわかる。なお、最終的な学習終了条件としては、目標回転速度から ±1度/秒 を許容している。また、Fig.3、Fig.4 の (d) を見ると、両者とも Actor 出力によって 2 つのロッドが 120 °の位相差を保って滑らかに変化していることがわかる。



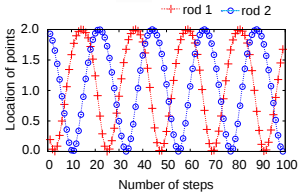
(a) Actor Outputs



(b) Rotation angle

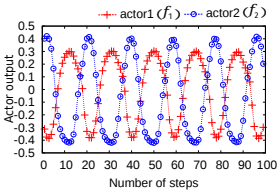


(c) Rotation speed

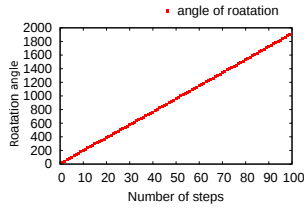


(d) Location of effort points

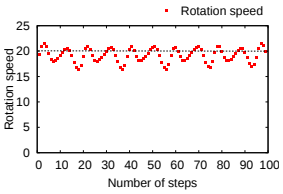
Fig. 3: System behavior in the test phase (desired rotation speed: 15 $degree/s$)



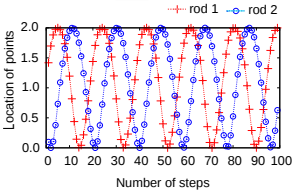
(a) Actor Outputs



(b) Rotation angle

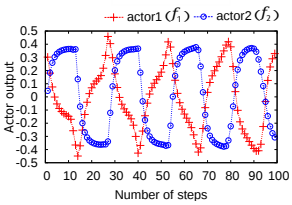


(c) Rotation speed

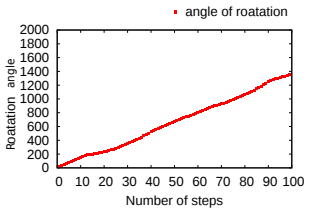


(d) Location of effort points

Fig. 4: System behavior in the test phase (desired rotation speed: 20 $degree/s$)

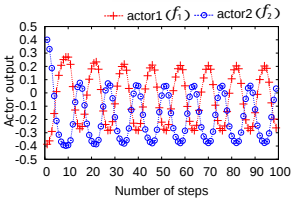


(a) Actor Outputs

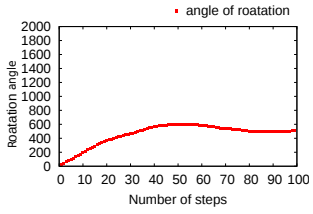


(b) Rotation angle

Fig. 5: Actor outputs and rotation when the inputs were cut off (desired rotation speed: 15 $degree/s$)



(a) Actor Outputs



(b) Rotation angle

Fig. 6: Actor outputs and rotation when the inputs were cut off (desired rotation speed: 20 $degree/s$)

次に、テストを 500step 続けた後、入力をその時の値に固定し、Actor の出力と車輪の回転角度を観察した。すると、目標回転速度 10度/秒 の時は、Actor の出力は振動しなくなったが、15、20度/秒 の場合は Fig.5、6 の (a) のように、入力の値が変化しないにもかかわらず、Actor の出力は Fig.5 の場合は入力を固定する前と近い形で、Fig.6 の場合は入力固定前より周期振幅ともに小さくなっているが、振動は継続した。このことから、強化学習を通して RNN 内部に振動子の機能が創発したと言える。一方、車輪の回転を見ると、Fig.5(b) では回転速度が落ちているものの回転を継続しているが、Fig.6(b) では回転は止まっている。これは Actor 出力の周期が小さくなっている一方で、車輪がそれに合わせた回転速度を維持するだけの Actor の出力の大きさがなかったためと考えられる。

最後に学習時の目標回転速度とテスト時の入力を固定する前と後の回転速度との間に相関関係があるかどうかを調べた。Fig.7 では、学習時の目標回転速度に対するテスト時の車輪の回転速度を△で、入力を固定した後に車輪が回転し続けた時の回転周期を○で示す。また、入力を固定した後に車輪が回転せずに神経振動子のみが振動する際は、その振動によって車輪が回転したと仮定し、その時の振動周期を回転速度に換算して□で示している。

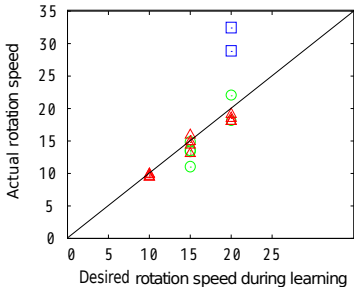


Fig. 7: The correlation of rotation speed in between learned oscillator and task requirement. The diagonal line represents that the desired rotation speed is identical to the actual rotation speed in the test phase

Fig.7 より、テスト時、入力を固定する前 (△) では、どの回転速度もほぼ線の上に乗っていることから、学習時の目標回転速度とほとんど同じ値を取っていることがわかる。そして、入力を固定した後では車輪を回転させ続けた時 (○) の回転速度は、学習した周期運動と近い回転速度となった。この時、入力は変化しておらず、タスクで求められる周期に近い周期の振動子が RNN 内部に創発したと言える。また、車輪を回転させることはできず、振動のみした場合 (□)、固定する前より振動が速くなっていることがわかった。一方、目標回転速度が 10度/秒 の時はどのパターンも神経振動子は形成されなかった。

5 まとめ

本研究ではピストン駆動方式の車輪を目標回転速度で回転させるタスクを RNN を用いた強化学習により自律学習させた。その結果、RNN がピストンを押す力を周期的に変化させ、車輪を目標回転速度付近で回転させることができるようになった。また、入力を固定することで RNN 内部に神経振動子の機能が創発されたかどうか調べたところ、学習した周期運動の回転速度が 10度/秒 では創発されなかったが、それより速い 15度/秒、20度/秒 の時には創発した。また、その際の Actor の振動周期は、一部例外があるものの、学習時に目標とした車輪の回転周期とほぼ一致した。このことから、タスクに応じた振動子を獲得できたと考えている。

最後に、本タスクの問題点として学習時に時間がかかり過ぎることが挙げられる。これに対しては何らかの対策が必要である。さらに、今回学習された振動子の周期のレンジが狭いため、その理由を解析するとともに、時定数の変化を学習によって実現し、レンジを広げることを検討していきたい。また、本研究において、タスクの学習はできるものの、振動子が形成される場合とされない場合の違いがどうして生じるのかについても解析していきたい。また、文献 [4] を参考に、状況に応じて周期を可変にする振動子が創発するかどうかも確認していきたい。

参考文献

[1] 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, Vol48, No.1, 計測自動制御学会, pp.106-111 (2009)

[2] 琴坂信哉, Stephan Schaal, ”神経振動子を用いたロボットのリズミックな運動生成”, ロボット学会誌, Vol.19, No.1, pp.116-123 (2001)

[3] J.Tani, Y.Yamashita, ”Emergence of Functional Hierarchy in a Multiple Timescale Neural Network Model:A Humanoid Robot Experiment”, PLoS Computational Biology, Vol.4, pp.2-6 (2008)

[4] J.Tani, ”Self-Organization of Behavioral Primitives as Multiple Attractor Dynamics: A Robot Experiment”, IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans, Vol.33, No.4, pp.481-488 (2002)