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

○沢津橋由人　Mohamad Faizal Bin Samusudin　柴田克成　（大分大学）

Emergence of Discrete and Abstract State Representation through Reinforcement Learning with a Recrrent Neural Network

∗Yoshito Sawatsubashi and Mohamad Faizal Bin Samusudin and Katsunari Shibata
(University of Oita)

Abstract− “Concept” is a kind of discrete and abstract state representation, and is considered useful for efficient action planning. However, it is supposed to emerge in our brain as a parallel processing and learning system through the learning based on a variety of experiences, and so it is difficult to be developed by hand-coding. In this paper, as a previous step of the “concept formation”, it is investigated whether the discrete and abstract state representation in formed or not through learning in a task with multi-step state transitions using Actor-Q learning method and a recurrent neural network. After learning, an agent repeated the sequence two times, in which it pushed a button to open a door and moved to the next room, and finally arrived at the third room to get a reward. After the learning, discrete and abstract state representation not depending on the door opening pattern was observed in two hidden neurons. The result of the learning with two recurrent neural networks that are for Q-values and for Actor outputs suggested that the state representation emerged to generate appropriate Q-values.

Key Words: Recurrent Neural Network, Reinforce Learning, Discrete State Representation

1　まえがき

われわれ人間は目や耳などを有し、時々刻々と膨大なセンサ信号が入力され続ける中で的確に状況を判断し、適切に行動することができる。我々が行動計画を立てる時は、個々のセンサ信号を意識するのではなく、例えば「部屋」とか「廊下」といった抽象的な表現で状態を表現し、「扉を開けて部屋を出て、廊下を出る」といったように行動の計画を立てる。このような離散的かつ抽象的な状態表現が、われわれの効率的な学習を実現しているとともに、高い知能の源ではないかと考えることができる。

では、なぜわれわれ人間はどんな「部屋」にいても、ここは「部屋」だと認識できるようになるのだろうか。部屋の中にいても壁の色が違えば全く違う視覚センサ入力になるであろうし、通常目で見えるものは部屋の一部分であり、目や頭や体を動かすと見える場所が変化してセンサ信号は大きく変化する。そう考えると、単に時系列データとしての膨大なセンサ信号を見ているだけで「部屋」という状態の切り分けができるようになるとはとても考えられない。部屋の中では、仕事をしたり、休んだりし、移動したい時はドアを開け、暑いときは窓を開けるといった行動をする。その時に、壁の色が何であるかよりも「部屋」にいるのか「廊下」にいるのか、はたまた「屋外」にいるのかの区別を付けた方が有効であるため、行動の学習の積み重ねを通してセンサからの膨大な時系列データから、さまざまな情報を複雑に並列処理し、「部屋」という認識ができるようになるのだと思われる。このような離散的かつ抽象的な表現をここでは “概念” と呼ぶ。

そこで “概念” といった高次機能をロボットにも与えることができれば、人間のように実環境でも柔軟な対応が可能になるのではないかと期待できる。しかし、単に窓やドアがあれば「部屋」というわけではなく、個々の “概念” は脳による複雑な並列処理によって形作られているため、これを人間の手で設計して与えることは不可能に近い。したがって、ロボットに “概念” を与えようとするなら、行動の学習を通して脳のような並列処理装置の中に創発させるアプローチが求められる。そして、“概念” が離散的かつ抽象的な状態表現であるとするならば、まず、行動の学習を通してそのような状態表現および状態間の遷移を表現できるようになることが必要であろう。

学習によって時系列の入力データから抽象的な表現を獲得する過去の方法として、谷らの研究が挙げられる [1)2)]。彼らはリカレントニューラルネット (RNN) を用いたモジュール回路群や複数の時定数の異なる部分回路から構成される RNN を導入し、現在のセンサ信号と動作から次の時刻のセンサ信号の予測を学習することにより、意味を持つ表現の分節化が発現すると報告している。しかし、センサ信号が視覚信号のように高次元で膨大な情報を持つ場合、時々刻々と入る入力を全て予測することは困難であるとともに、前述のように、行動生成に対する必要性を考慮せずセンサ信号だけから「部屋」という状態を切り分けることは無理ではないかと考えられる。またここで獲得された抽象表現は連続的な変化をしており、離散的な状態表現にはなっていない。

行動の学習を通して並列処理装置の中に抽象的な状態表現を創発させるためには、ニューラルネット (NN) と強化学習の組み合わせが有効であると考えられる。強化学習は、報酬を得て罰を避けるという目的に沿った行動を自律的に学習させることができる。その強化学習に情報を並列処理し学習できる NN を組み合わせれば、NN が強化学習に基づいて最適化されることで、適切な行動を実現するために必要な情報が浮かび上がり、抽象的な表現を獲得することが期待できる。また状態遷移タスクを学習することで、分節化された状態

として離散的な抽象表現とその遷移が NN 内部に表現されるようになることが期待される。従来、筆者らのグループで行った研究では、入力自体に 2 値の信号が入っていたり、状態遷移が 1 回のみの場合について、学習によって離散的な状態遷移を表現できるようになることを確認している [3)4)]。

そこで本研究では、概念獲得の前段階として、シミュレーションによって多段階の状態遷移が必要な複数の部屋とドアからなる環境のタスクにおいて、NN と強化学習を組み合わせて学習することで、状態の切り替わりを表す 2 値の信号の入力なしで NN 内部に目的に沿った離散的な抽象表現が創発するかを観察した。

2 学習方法

本研究では、学習システムとして Fig.1 のような 5 層の Elman 型のリカレントニューラルネット (RNN) と強化学習の組み合わせを用いる。ここでは、「スイッチを押す」「移動する」などのように離散的な選択を求められるものを「行動」と呼び、具体的に x 方向、y 方向にどれくらい移動するかのように連続値としての動きを求められるものを「動作」と呼んで区別し、両者を併用して agent の動きを決定する。そこで、離散的な行動選択が可能な Q-learning と連続値出力が可能な Actor-Critic を組み合わせた Actor-Q learning を用いた [5)]。そのため RNN の出力は離散行動のための Q 値と連続動作を決定する Actor の 2 種類を用意した。そして、まず離散的な行動を Q 値の出力を基に選択し、選択された行動が連続値としての動作を必要とする場合に Actor の値を用いる。そのため、行動によっては動作である Actor を用いない場合もある。

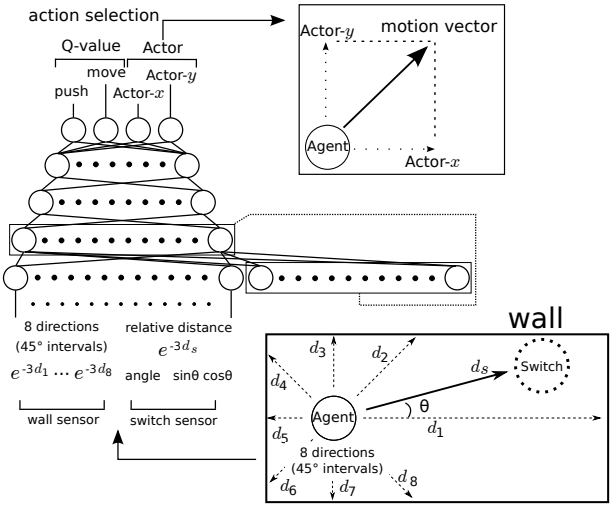


Fig. 1: Learning system consisted of a 5-layer Elman-type RNN, and its inputs and outputs.

学習は、強化学習の Q-learning と Actor-Critic のアルゴリズムを基に RNN の教師信号を自動生成し、Back Propagation Through Time(BPTT) 法を用いて RNN 内の結合重み値を更新していく [6)]。Q 値の更新は、現在の状態 $\mathbf{s_t}$ での行動 a_t に対する Q 値 $Q_{a_t}(\mathbf{s}_t)$ と次 step の状態 $\mathbf{s_{t+1}}$ での Q 値の最大値 $\max_a Q_a(\mathbf{s}_{t+1})$ を用いた TD 誤差 $TDerror_t$ により更新を行う。時刻 t の行動 a_t に対する Q 値出力の教師信号 $Q_{d\ a_t,t}$ を以下の式に示す。

$$\begin{aligned} Q_{d\ a_t,t} &= Q_{a_t}(\mathbf{s}_t) + TDerror_t \\ &= r_{t+1} + \gamma \max_a Q_a(\mathbf{s}_{t+1}) \end{aligned} \quad (1)$$

$$TDerror_t = r_{t+1} + \gamma \max_a Q_a(\mathbf{s}_{t+1}) - Q_{a_t}(\mathbf{s}_t) \quad (2)$$

ここで γ は割引率、r_t は報酬である。選択された行動以外の Q 値の出力は学習しない。また、行動 a_t が「動作」を伴う場合は Actor の学習を行う。Actor は、Q 値に基づいて生成された TD 誤差 $TDerror_t$ を用いて更新を行う。Actor の教師信号 $\mathbf{A}_{r,t}$ を以下の式に示す。

$$\mathbf{A}_{r,t} = \mathbf{A}(\mathbf{s}_t) + TDerror_t \times \mathbf{rnd}_t \quad (3)$$

ここで $\mathbf{A}(\mathbf{s}_t)$ は Actor の出力である動作ベクトル、$\mathbf{rnd}$ は実際に動作する際に探索のために動作ベクトルに加えた乱数ベクトルである。

3 タスク設定

Fig.2 に隣接した複数の部屋を想定した状態遷移タスクの環境を示す。agent は、最初 Fig.2 の中心の部屋である first room 内に配置される。agent は部屋の中心に配置してあるスイッチを押すことで、現在いる部屋と隣の部屋の second room を結ぶドアの 1 つがランダムに開けられる。そして、second room に移動すると、その中心にも同じようにスイッチがあり、それを押してランダムに開いたドアの先にある最終目的の部屋である third room へと 2 部屋移動するタスクである。third room までの経路は Fig.2 の矢印が示すように 12 パターン用意した。

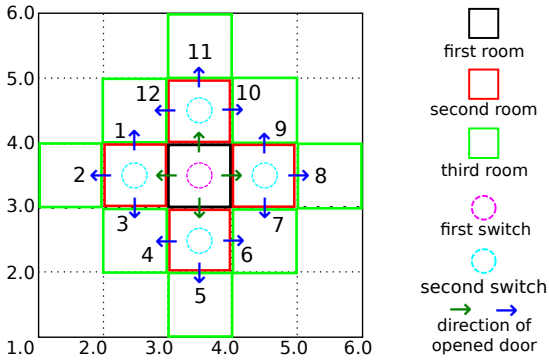


Fig. 2: Task environment with rooms and switches.

Fig.1 のように、agent はセンサ信号として部屋の壁までの距離情報を 45 °間隔で $d_1 \sim d_8$ の 8 個、スイッチの位置情報を距離 d_s と方向 $sin\theta$ と $cos\theta$ の 3 個の計 11 個で与えられる。このとき距離が近くなるにつれ入力を大きくするために距離 $d_i(i = 1, \cdots, 8, s)$ は e^{-3d_i} として入力した。また、RNN に入力するスイッチの情報は、Fig.2 のように first switch と second switch のように複数ある場合は、押した押さないに関わらず agent と距離が近い方、つまり、自分がいる部屋にあるスイッチの情報とした。

このとき agent には入力が与えられるが、入力の値が何を意味しているかわからないため、agent は報酬を基に行動を学習する中で、入力の意味を把握しなければならない。ドアが開いて新しい部屋が出現したことは agent に直接与えていないため、8 方向の測距センサ

の値から判断する必要がある。また、正しい進行方向を知るためには、自分がどちらから来たかを覚えておかなければならないため記憶が必要なタスクとなる。

出力は 4 つ用意し、「移動」と「スイッチを押す」という 2 つの行動の Q 値と、「移動」選択時の移動量を決定する x 軸方向と y 軸方向の Actor に用いた。agent の 1 回の移動は現在の位置から半径 0.5 の円形の範囲内とした。学習の終了条件は、agent が 35step 以内に 2 部屋移動し、報酬を得ることを 3 万回連続で成功した場合とした。また 1 試行で 200step 経過しても報酬を得られない場合は失敗とみなして、そこで 1 試行を打ち切った。行動選択は Bolzmann 選択を用い、学習が進むにつれ温度を 1 から 0.01 へと徐々に下げていった。その他の設定は Table 1 に示す。

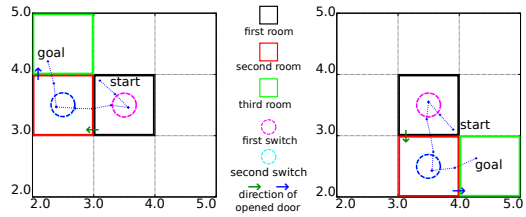
Table 1: Parameter setting.

initial position of the agent (x, y)	
(3.0∼4.0, 3.0∼4.0)	
radius of the switch	0.2
number of layers	5
number of neurons in each layer	
11(input)-80-40-20-4(output)	
constant input for bias	0.1
value range of sigmoid function	-0.5∼0.5
reward r at the third room	0.4
punishment for "push" action at the outside of the switch	-0.1
discount factor γ	0.9
range of the exploration **rnd** (rnd_x,rnd_y) for actor	
-0.5∼0.5 → -0.2∼0.2	
traced back time in BPTT	30
initial connection weight	
input - hidden1	-0.1∼0.1
hidden1 - hidden2	-0.2∼0.2
hidden2 - hidden3	-0.5∼0.5
hidden3 - output	0.0
self-feedback	4.0
non self-feedback	0.0
learning rate	
for feedback connections	0.0125
for other connections	0.5

4 学習結果

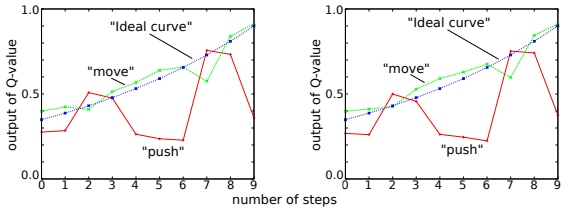
乱数系列を変化させて 5 回学習させたところ、5 回とも学習は成功し、学習終了までの平均試行回数は 765,486 試行であった。さらに学習後の RNN を用い、first room 内の agent の初期位置を x, y のそれぞれを 0.01 刻みで配置した全 9,801 通りを各部屋の出現パターン毎にテストを行った。その結果、agent は 9,801 通り全部屋の出現パターンにおいて目的の部屋にたどり着いた。しかし、学習に用いる乱数系列によっては、学習の終了条件は満たしてもテストにおいてわずかながら失敗する場合もあった。

Fig.3∼Fig.6 に学習後の agent の行動例を示す。Fig.3∼Fig.6 において、(a) は agent の初期位置 (3.1, 3.9) でドアが左、上の順で開いた場合、(b) は agent の初期位置 (3.9, 3.1) でドアが下、右の順で開いた場合を示している。(a),(b) それぞれ Fig.3 が agent の経路、Fig.4 が Q 値の出力、Fig.5 が Actor の出力、Fig.6 が最上位の中間層ニューロンの出力の一部の 1 試行中の時間変化を示している。



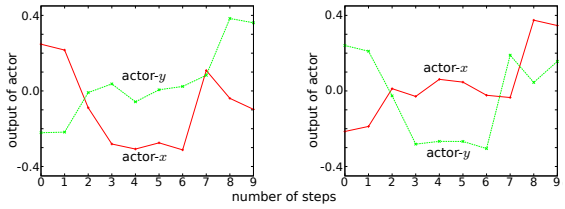
(a) (b)

Fig. 3: Two sample trajectories of the agent after learning.



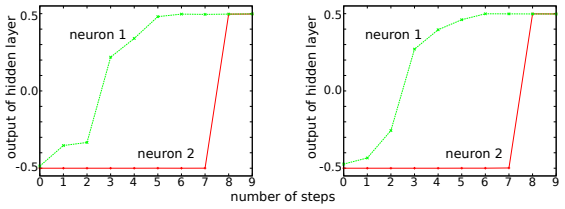
(a) (b)

Fig. 4: Change of the Q-value outputs in one episode.



(a) (b)

Fig. 5: Change of the Actor outputs in one episode.



(a) (b)

Fig. 6: Change of outputs of two hidden neurons in the Actor-Q-RNN in which the output changes discretely in one or a few steps.

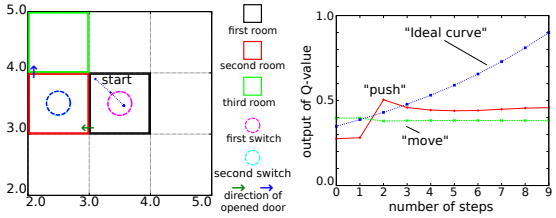
Fig.3 を見るといずれの場合も agent は各部屋のスイッチ上に移動し、スイッチを押し、目標の部屋に到達していることが分かる。2 つ目のスイッチを押した場合に、agent から見た部屋の配置は (a) と (b) で同じになるため、入力のセンサ信号だけでは新しく出現した部屋がどちらかわからない。しかし、それでも agent は迷うことなく目的の部屋に移動していることから、その時点でのセンサ信号だけでなく記憶に基づいて行動できていると言える。

Fig.4 を見ると、各 step で "move" と "push" の大きい方の Q 値が、ほぼ理想曲線と近い形で単調に増加している。agent がスイッチ上に位置する step2、step7 では、スイッチを押してドアを開けるように "push" の評価が増加している。このことから、agent はスイッチ上

にいることを正しく認識し、行動価値に反映させることができていると言える。Fig.5 では新しく出現したスイッチやドアに向かうように 2 つの Actor の出力が変化していることがわかる。

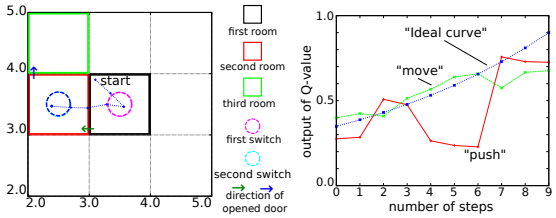
Fig.6 は最上位の中間層ニューロン 20 個のうち、特徴的な反応をする 2 つのニューロンの step ごとの出力変化を示す。neuron1 は、1 つ目のスイッチを押した後、1step で完全に値が反転しているわけではないが、出力が負から正に変化し、目的の部屋までその値を保持している。neuron2 は、2 つ目のスイッチを押した後、出力が負から正に変化し、目的の部屋まで出力を保持している。この 2 つの中間層ニューロンの出力の変化は、部屋の配置を変えてもあまり変化しなかった。

またこれらニューロンの役割を見るため、出力を固定し、どのような影響が出るか観察した。1 つ目のスイッチを押した後に出力が変化する neuron1 を負の値-0.5 に固定した場合、Fig.7(a) の agent の経路で示されるように、agent は 1 つ目のスイッチを押すことはできるものの、その後 second room に移動できなかった。Fig.7(b) の Q 値の変化から見ると、agent は 1 つ目のスイッチを押した後、移動へと行動の切り替えができていない。同様に neuron2 を負の値-0.5 に固定した場合、Fig.8 に示されるように 2 つ目のスイッチを押した後、移動へと行動の切り替えができなかった。これは neuron1,2 が出力を負から正へと変化することでドアが開いた状態かどうかを区別しているからだと考えられる。そのため、agent はこの neuron の変化がないためドアを開けても移動の評価値が上がらないと考えられる。



(a)A sample trajectory (b)Q-value change

Fig. 7: Learning result when the output of the hidden neuron1 was fixed to -0.5.



(a)A sample trajectory (b)Q-value change

Fig. 8: Learning result when the output of the hidden neuron2 was fixed to -0.5.

5 離散的な抽象表現の発現理由

なぜこのように、ドアが開いていない状態、1 つ目のドアが開いた状態、2 つ目のドアが開いた状態というどのドアが開いたかによらない離散的な状態の区別をする表現が獲得できたのだろうか。過去の研究では[3)4)]、スイッチを押しているという 2 値入力を入れていたため、それが表現獲得に大きく影響していたと考えられる。しかし本タスクでは、ドアを開けた場合に離散的な入力の変化はあるものの、測距センサの信号が 0 から 1 までの値域を持つ状況でおよそ 0.22 から 0.01 ほどの小さな変化しかない上、どちらのドアが開くかによって変化する入力が異なることから、単に値を記憶したということでは説明がつかない。おそらくタスクにおける必要性、つまり、強化学習による適切な状態評価と行動の学習によって離散的な抽象表現が創発したと考えられる。

そこでまず、どちらの出力の学習がこのような表現の創発を引き起こすのかを調べるために Fig.9 のように Q 値用と Actor 用の 2 つの RNN に分けて学習し、RNN の観察を行った。中間層ニューロン数は、入力層から 11-30-20-10-2 とした。

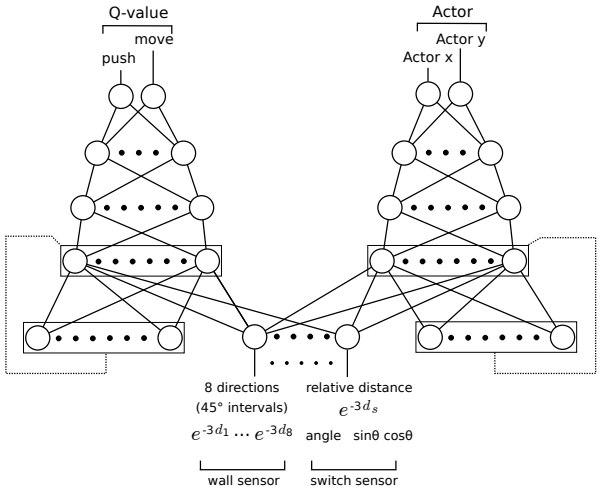
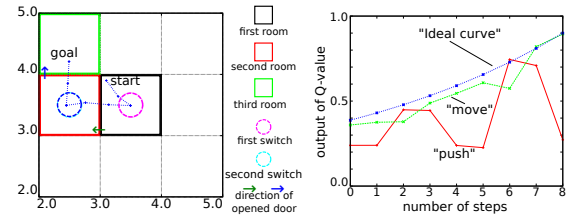


Fig. 9: Learning system when the RNN is divided to Q-RNN and Actor-RNN.

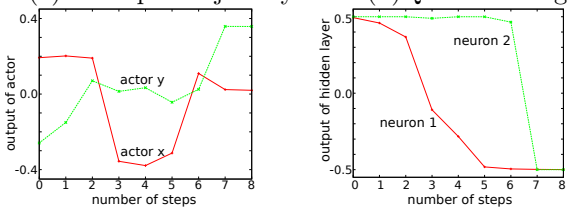
乱数系列 5 回分学習した結果、5 回とも学習が成功し、平均試行回数は 1,878,943 試行であった。学習後、前節と同様に agent の初期位置を変更し、各部屋の出現パターン毎にテストを行った。その結果、全てのパターンにおいて agent は目的の部屋にたどり着いた。

Fig.10 に学習結果を示す。Fig.10 の (d) のように、Q 値用の RNN の最上位の中間層で 1 つ目のスイッチを押した後に出力が離散的に変化するニューロン (neuron1) と、2 つ目のスイッチを押した後に出力が離散的に変化するニューロン (neuron2) が見つかった。このとき最上位の中間層である neuron1 と結合する出力層との重み値を確認すると、Q 値の “push” とは 0.004、“move” とは-0.911 であった。このことから neuron1 は出力が負になり “move” との重み値も負の値であることから move の Q 値を上げる働きをしていると分かる。また neuron2 と出力層との重み値を確認すると Q 値の “push” とは 0.426、“move” とは-0.472 であった。このことから neuron2 は出力が負になることで “push” の Q 値を下げ、“move” の Q 値を上げる働きをしていると分かる。

また、前節と同様に neuron の出力を固定して行動への影響を確認した。Fig.11 は neuron1 の出力を 0.5 に固定、Fig.12 は neuron2 の出力を 0.5 に固定した場合であり、(a) は agent の移動経路を (b) は Q 値の変化を示している。結果を見ると前節の場合と同じように、スイッチを押した後に行動を “移動” に切り換えることができず、報酬を得ることができなかった。この



(a)A sample trajectory (b)Q-value change



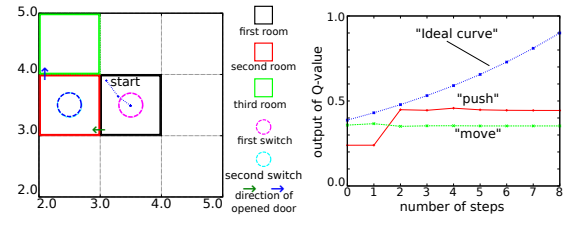
(c)actor change (d)Change of outputs of two hidden neurons in the Q-RNN

Fig. 10: Learning result when the RNN is divided to Q-RNN and Actor-RNN.

ことからも neuron1,2 は前節と同じような役割を持ったニューロンであり、状態が切り替わったことを表現するニューロンであることがわかる。

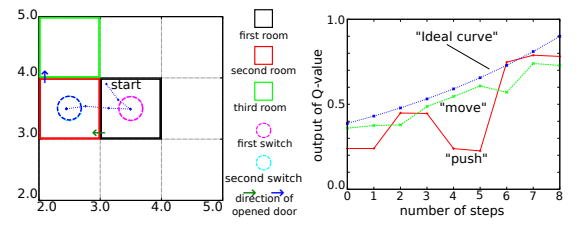
こういった開くドアによらない離散的な抽象表現は、Q 値側の RNN にしか確認できなかった。よって本タスクにおける離散的な抽象表現は適切な Q 値を実現するために創発したと考えられる。Q 値は、スイッチ上で “スイッチを押し” てドアを開け、次の部屋が現れたかどうか判断し、“移動する” へ行動が切り換わるように変化することが必要である。また、行動を大きく切り換えるためには、ドアが開いたという入力の小さな変化を読み取り、それによって Q 値出力を的確に変化させるために、中間層において入力の小さな変化を拡大して表現するように学習されたのではないかと考えられる。

また、Q 値は、ドアの開く方向の違いによって入力は違っていても、「移動」という同じ出力を変化させるために、中間層において入力の違いによらない表現をするように学習されたのではないかと考えられる。さらに、同じ行動に対する価値でも部屋が変わってゴールが近づくとその値を変化させる必要から first switch を押すと反応するニューロンと second switch を押すと反応するニューロンが別々に発現したと考えられる。



(a)A sample trajectory (b)Q-value change

Fig. 11: Learning result when the output of the hidden neuron1 was fixed to 0.5 in the Q-RNN.



(a)A sample trajectory (b)Q-value change

Fig. 12: Learning result when the output of the hidden neuron2 was fixed to 0.5 in the Q-RNN.

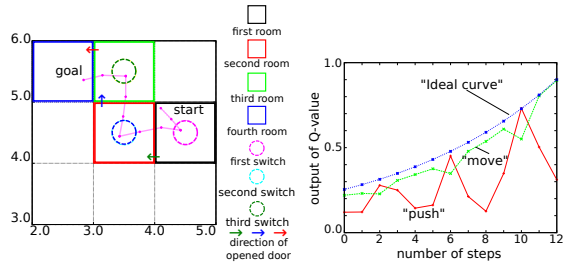
自律学習の観点から考えると、“スイッチを押す” と “移動” の行動を用意すること自体が大きな知識の付与となり、そもそも “スイッチを押す” と “移動” という行動の切り分けがどうしてできるようになるかという問題が残る。元々 “行動” というものはなく、アクチュエータレベルの “動作” のみで考えるとすると、“行動” は “動作” を大きく切り換えるための状態表現と考えることもできる。ドアが開く前には、“スイッチを押す” ための動作、ドアが開いた後では、移動方向は違っていても、車輪や足を使って “移動する” という類似した動作へと動作の切り換えを表現するために、離散表現が獲得される可能性があると考えられる。

このことから強化学習における報酬に基づいた自律学習が離散的な表現を獲得していく上で重要であると考えられる。

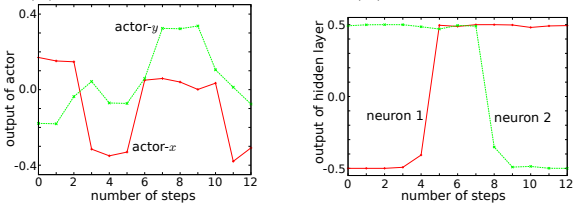
6 4 部屋移動タスク

さらに部屋とスイッチを追加し、4 部屋とスイッチ 3 個の環境に変更して、より多段階の状態遷移が必要なタスクの学習を行った。しかし、最初から 4 部屋目で報酬を与える設定にしたところ、agent は探索によって報酬を得ることが少なく学習が進まなかった。そのため、ここでは、最初から少ない部屋数から学習し、正しい行動を学習してから、部屋数を増やしていくように学習方法を変更した。学習の最終終了条件は、1 千万試行までに 3 万回連続で 75step 以内に報酬を得た場合とした。

乱数系列 10 回分学習した結果、終了条件を満たしたものは 4 つあり、4 つの平均試行回数は 7,194,717 試行であった。Fig.13 は、このうちの 1 回の学習結果を示す。



(a)A sample trajectory. (b)Q-value change.



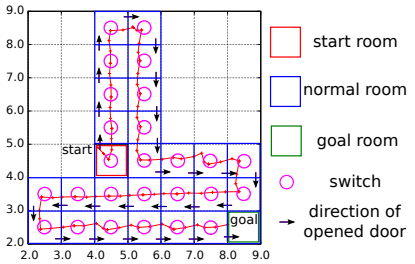
(c)actor change. (d)Change of outputs of two middle hidden neurons.

Fig. 13: Learning result in the four-room task.

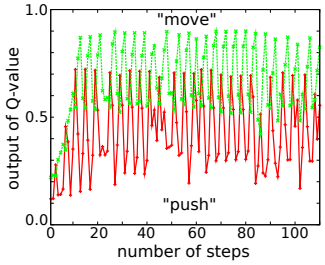
前述の 3 部屋の移動タスクでは、最上位の中間層ニューロンにて、離散的な状態を表現するニューロンが観

察されたが、ここでは、真ん中の層において、Fig.13 の (d) のような離散的な状態表現する 2 つのニューロンが観察された。また、発火タイミングについても、3 部屋のタスクではスイッチを押して、新しい部屋が現れるタイミングで値が変化していたが、ここでは neuron1 は、移動して最初に部屋が変わるタイミングで、neuron2 は、2 つ目の新しい部屋が現れて 1step おいてから値が変化した。また、3 つ目の部屋が現れてから値が大きく変化するニューロンは見つからなかった。タスクの学習方法が違うことも、この表現の違いに影響している可能性もあるが、詳しい解析はまだできていない。

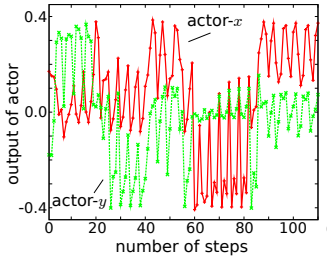
また、重み値を使用し、さらに部屋とスイッチの数を増やした環境に置いたところ、Fig.14 の (a) のように、学習していないにもかかわらず、26 個のスイッチを押して 27 個目の部屋まで到達することが確認できた。Q 値の変化を見ると、学習時に報酬を得ていた 4 部屋目に入ると移動の Q 値が下降し、その後、両 Q 値は 3 部屋目と類似した変化をして、再びスイッチを押してドアを開いた。多少の変化はあるものの、基本的にはこの流れを繰り返して 27 個目の部屋まで到達している。一方、Actor の方は、その都度新しく出現した部屋の方向に向かうように値が切り換わっていることがわかる。これらの結果は、「部屋」単位で行動、動作の汎化が起こっているように見え、「部屋」という単位での状態の切り分け、ひいては「部屋」という概念形成の可能性を感じさせる結果であると考えている。



(a)Sample trajectories of the agent after learning.



(b)Change of the Q-value outputs in one episode.



(c)Change of Actor outputs in one episode.

Fig. 14: Test result when the Actor-Q-RNN was used the weight value.

7 まとめ

本研究では，概念形成の前段階と位置づけ，多数の部屋とスイッチからなる環境でゴールへ向かう行動をリカレントネットを用いた強化学習で学習させることで，状態遷移を直接表す 2 値入力なしで，中間層ニューロンが多段階の離散的かつ抽象的な状態表現を学習できることを示した。そして，このような離散的かつ抽象的な状態表現が発現するのは，入力に違いがあっても類似した行動の切り替えを求められるためとの考えを示した。

「部屋」という概念の創発を示すためには,「部屋」と「廊下」といった違う状況が混在する環境で学習を行う必要があるとともに，視覚センサ信号のようなより多次元の情報から必要な情報を抽出し，そこから複合的に「部屋」であると判断できるようになることが重要であると考えられる。

しかし、本研究で創発した離散的な抽象表現は状態変化を捉えられているが、概念というにはまだ十分ではない。

概念を形成するためには、今回のタスクのように限られた情報を用いた学習ではなく、多数の入力を用いて学習や経験を積ませることが必要ではないかと考える。限られた情報による学習は、様々な事を考慮した抽象表現が創発されないために、情報に融通が利かない。そうではなく概念は多数の入力からの不要な情報を排除し、様々な事を考慮した抽象表現により複合的に判断することで創発するものだと考える。そのため本研究のような距離のみのセンサではなく様々な環境においても汎用的に使え、情報に縛りが少ない視覚センサなどを用いることでより情報を考慮した表現が学習により獲得できるため、概念形成に必要と考える。

しかし、本研究で確認した離散的な抽象化表現は概念獲得の足がかりになると考えられ、また概念獲得に向けて NN と強化学習による学習の有効性をある程度示せたのではないかと考える。

参考文献

1) Tani, J. & Nolfi, S. (1999) Learning to Perceive the World as Articulated: An Approach for Hierarchical Learning in Sensory-Motor Systems, Neural Networks, **12**, pp. 1131-1141

2) Y. Yamashita and J. Tani (2008) Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment, PLoS Computational Biology, **4**(11), e100220

3) M, F, Samsudin. & Shibata, K. (2012), Emergence of Multi-Step Discrete State Transition through Reinforcement Learning with a Recurrent Neural Network, Proc. of ICONIP 2012 (to appear)

4) Utsunomiya, H. & Shibata, K. (2009). Contextual Behaviors and Internal Representations Acquired by Reinforcement Learning with a Recurrent Neural Network in a Continuous State and Action Space Task, Advances in Neuro-Information Processing, Lecture Notes in Computer Science, Vol.5507, pp. 970-978, 5507-0970. pdf (CD-ROM)

5) Shibata. K, Tetsuo Nishino & Yoichi Okabe (2002) Active Perception and Recognition Learning System Based on Actor-Q Architecture Systems and Computers in Japan, **33**(14), pp. 12-22

6) Rumelhart, D.E, Hinton, G.E., and Williams, R.J.: Learning Internal Representations by Error Propagation, Parallel Distributed Processing, The MIT Press, pp. 318-362 (1986)