社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

リカレントネットを用いた強化学習による探索行動と多値記憶の創発

柴田　克成†　　後藤　健太†,††

† 大分大学工学部電気電子工学科
大分市大字旦野原 700 番地
†† 現在，ナブテスコ (株) 勤務
E-mail: †shibata@oita-u.ac.jp

あらまし　著者らは，ニューラルネットを用いた強化学習を行うことでさまざまな機能が合目的的かつ調和的に創発することを提唱してきた．本稿では，記憶を利用した決定論的な探索行動の創発に焦点を当てる．ゴールが見えない 3x3 のランダム迷路タスクの Q 学習において，リカレントネットを導入することでエージェントは過去の経験を考慮したより適切な Q 値を表現し，より適切な探索行動を学習することができること，さらに，未知の環境でもある程度有効に働くことを確認した．また，分岐位置がランダムに出現する単純な環境での学習では，最適行動実現に必ずしも必要ではないが，適切な Q 値を表現するために多値の分岐位置を記憶するようになることを示した．

キーワード　強化学習，リカレントニューラルネット，探索の学習，記憶，機能創発

Emergence of Exploration Behavior and Multi-valued Memory through Reinforcement Learning with a Recurrent Neural Network

Katsunari SHIBATA† and Kenta GOTO†,††

† Oita University
700 Dannoharu, Oita, JAPAN
†† Currently, Nabtesco Corporation
E-mail: †shibata@oita-u.ac.jp

Abstract　The authors have propounded that various functions emerge purposively and harmoniously through reinforcement learning with a neural network. In this paper, emergence of deterministic “exploration” behavior utilizing memory is focused on. In the simulation of 3 × 3 random maze with an invisible goal task, by introducing a recurrent neural network for Q-learning, an agent could represent more accurate Q-values considering past experiences, and learn more appropriate exploration behaviors. The acquired knowledge could be generalized in some unknown environments to some extent. It is also shown that through the learning in a simple environment with a random-located branch, the recurrent neural network memorizes and keeps the multi-valued branch position to represent accurate Q-values even though that is not required to realize the optimal path.

Key words　reinforcement learning, recurrent neural network, learning of exploration, memory, function emergence

1. はじめに

知能ロボットの研究は長年行われて来ているものの，柔軟性は未だ人間にはるかに及ばない．両者の処理を見ると，われわれの脳は超並列で柔軟であり，一方ロボットでは，通常，設計者としての人間によって開発された機能モジュールが直列に接続されて構成されている．認識や制御と異なり，高次機能はセンサやアクチュエータと直接つながっておらず，何を入力とし，何を出力とするかすら決めることが困難であるし，それを予め与えてしまうと今度は柔軟性が失われる．このような観点から著者らは，人間のような柔軟性や知能をロボットに実現するためには，その構成や学習方法をより人間に倣わなければならないと考えている．そして，ニューラルネットを用いた強化学習を通してさまざまな機能を合目的的かつ調和的に創発させることを提唱し，個々の機能が実際に創発するかどうかを調べてきた [1] [2]．その中の一つが，「決定論的探索」である．

強化学習の分野で「探索」と言えば，通常，乱数を用いた確率的な行動選択を指す．効率的な学習のために，各状態の出現

回数を記録するなど過去の行動を考慮した能動的な探索は研究されてきた [3] [4]．しかし，われわれ人間は鍵をなくすと，鍵がある可能性が高いと思うところや近いところから探していくなど状況に応じた柔軟かつ戦略的な探索を行うことができる．このような知的な探索のためには，現在の状況や過去の履歴を含むさまざまなことを並列に考慮する必要がある．そこで，著者らはリカレントニューラルネットを用いた強化学習を通した知的探索の創発の可能性を探り，簡単ではあるが，戦略的な行動の創発を示してきた [5] [6] [7].

知的探索の実現のために，リカレントネットがどのように必要な情報を記憶し，適切な探索行動に反映させるかが鍵となる．文献 [5] では, 本物と偽物の２つのゴールの目印があるマス目の環境で，エージェントは探索行動を学習した．エージェントは，片方の目印に最初に行き，そして，それが本物でない場合はもう一つの目印に行くという探索行動を獲得した．さらに，どちらかの目印をすでに訪問したかどうかを表現する中間層ニューロンが創発することを示した．

本稿では，ゴールの目印もなく，エージェントは回転して向きが変わるより難しいタスクにおいて，常時入力されるセンサ信号から，すでに通ったマスであるかどうか，分岐位置はどこだったかというような探索に役立つ情報を抽出して記憶し，効率的に探索できるようになることを示す [8].

また記憶の創発の観点からは，すでに，リカレントネットを用いた強化学習によって，後の行動選択に必要となる２値の情報の記憶が創発することが示されている [9] [10]．文献 [11] [12] では，予測を必要とするタスクにおいて，学習を通して２値でない多値の情報を中間層ニューロン間での情報のリレーによって記憶するようになったことを示した．本稿ではこれらと同様に，何を記憶すべきかを明示せず，強化学習を通した多値の情報を記憶する能力の創発を示すことも目的とする．

2. 学習システム

学習システムは非常に簡単なもので，図 1 に示すように，一つのリカレントニューラルネットがあり，そこにエージェントが知覚した信号が入力される．リカレントネットとして良く使われる 3 層の Elman ネット (中間層ニューロンの出力が次の時刻の入力として与えられるもの) を用いて，離散時間モデル [13] として計算を行った．本稿では，離散空間タスクを扱うので，強化学習のアルゴリズムは Q 学習 [14] を用いた．リカレントネットの出力の数はエージェントの行動の数と等しい．ここでは，各ニューロンの出力関数として-0.5 から 0.5 の値域のシグモイド関数を用い，リカレントネットの出力に 0.4 を足したものを Q 値として用い，逆に，Q 値の理想値から 0.4 を引いてリカレントネットの教師信号として用いた．行動選択にはボルツマン選択 [15] を用いた．実行した行動 a_t に対応する出力のみ，Q 学習に基づいて教師信号 $T_{a_t,t}$ を

$$T_{a_t,t} = r_{t+1} + \gamma \max_{a'} Q_{a'}(\mathbf{S}_{t+1}) \qquad (1)$$

と生成して学習した．ただし，r は報酬，γ は割引率，Q_a は行動 a に対する Q 値，$\mathbf{S}$ はネットワークへの入力ベクトルを表す．ネットワークは BPTT (Back Propagation Through Time) [13] で教師あり学習を行う．このように，システムも学習も非常に簡単で一般的なものであり，探索の学習のための特別な手法は一切用いていないことに注目して頂きたい．

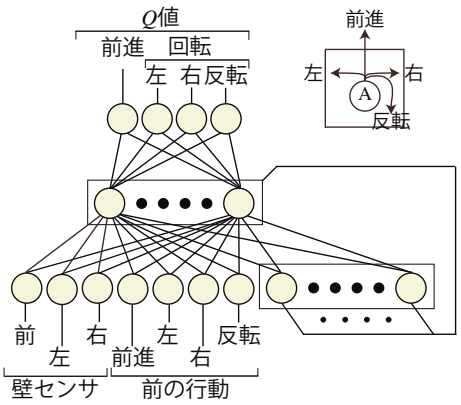

図 1　エルマンネットで構成した学習システムと入出力

Fig. 1　Learning system consisted of one Elman-type recurrent neural network and its inputs and outputs.

3. シミュレーション

本稿では，試行ごとに変化する離散の環境で，ランダムに置かれる見えないゴールをエージェントが探すタスクを扱う．まず，3 × 3 の迷路タスクで学習をさせ，次に，記憶機能の獲得をより詳しく調べるため，一つの分岐を持つより単純な環境で学習させる．

3.1 タスク設定

3 × 3 の迷路タスクでは，図 2 のように，9 個のマスを 3 × 3 の形に配置した．エージェントは常に真ん中のマスからスタートし，ゴールは中央以外の 8 個のマスにランダムに置き，試行中は固定した．外部との境界である 12 個の壁以外に，エージェントがすべてのマスに移動できる条件で毎試行 4 個の壁をランダムに配置した．壁の配置パターンは全部で 192 通り，ゴール位置との組み合わせは全部で 1536 通りである．エージェントは,「前進」「右回転」「左回転」「反転」の 4 つの行動からボルツマン選択によって確率的に選択する．エージェントは，すぐ前，左，右の 3 か所のそれぞれに壁があるかないかの局所的な情報を示す 3 つの信号と直前に行った行動が何かを表す 4 つの信号の合わせて 7 個の 2 値の信号をニューラルネットに入力する．ボルツマン選択はニューラルネットの外で行われるので，実際に行った行動を把握するため，直前にとった行動もニューラルネットに与えた．エージェントが見えないゴールのマスに到達すると報酬が得られ，その試行は終了する．また，壁にぶつかっても罰は与えなかった．今回は，エージェントは「回転」の行動により向きを変えるので，同じく 3 × 3 の迷路タスクを扱った [5] や [7] よりも状況の区別が難しいタスクとなっている．

このタスクでは試行の最終ステップで一定の報酬が与えられるため，エージェントは，ゴールに至るまでのステップ数の期待値が小さいほど Q 値が大きくなるように学習し，Q 値が大きい行動をより大きい確率で選ぶ．ゴールまでのステップ数の

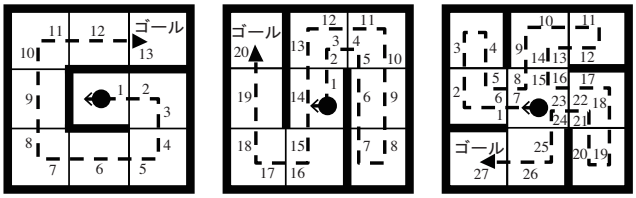

図 2　3 つのサンプル壁配置の場合における学習後のエージェントの行動．小さな数字は試行中の何ステップ目かを表す．

Fig. 2　Sample agent's behaviors after learning for three types of wall allocations. The small numbers indicate the step number in each episode.

期待値は，これまでの行動履歴 (文脈) を考慮する必要があるため，学習を通して，必要な記憶を形成することが期待される．他のパラメータ設定を表 1 に示す．

表 1　各パラメータの設定

Table 1　Parameter Settings

学習した試行回数	500,000
ステップ数上限	250
ボルツマン選択での温度	0.05 → 0.0025
割引率 γ	0.92
ゴール時の報酬 r	0.8
外部入力の数	7
中間層ニューロン数	40
出力層ニューロン数	4
中間層ニューロンの初期値	0.0
BPTT でさかのぼるステップ数	30
初期重み値	
中間層 → 出力層	0.0
外部入力 → 中間層	random from -0.5 to 0.5
セルフフィードバック	4.0
その他のフィードバック	0.0
学習係数	
フィードバック結合 (3×3 タスク時)	0.1 → 0.05
その他の結合 (3×3 タスク時)	0.2 → 0.1
すべての結合 (分岐タスク)	0.5 → 0.25

3.2　3 × 3 迷路問題

まず，3×3 の迷路問題を学習させた．図 3 に，1,000 試行ごとにゴール到達までの平均ステップ数を示した学習曲線を示す．図 2 は，壁配置の 3 つのサンプルの場合の学習後のエージェントの行動を表す．テスト時は最大 Q 値にしたがってエージェントを行動させ，ゴールは配置せずに最後のマスに至るまでその行動を観察した．ここでは，エージェントは最初に左を向いているとした．図 4 は，図 2 の場合の各ステップでエージェントが選択した行動の Q 値，つまり，最大 Q 値の変化を示す．プロット点のない線は理想的な Q 値の変化を示す．

すべての場合で Q 値は理想に近い値となっている．環境 1 では，エージェントが向きを変えながら道なりに進んでいくと，次のマスがゴールである確率が最初は 1/8 = 12.5% であったものが，最後のマスに行く際には 100%と徐々に高くなる．学

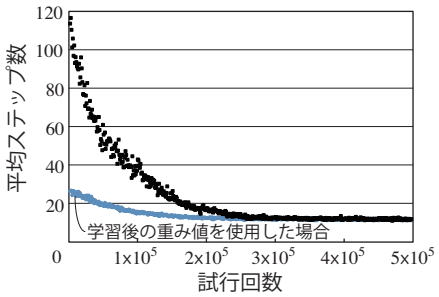

図 3　3 × 3 の迷路タスクでの学習曲線．このプロットはランダム探索の影響を含むので，学習後の重み値で固定した場合のランダム探索によるステップ数の変化を比較のためにプロットした．

Fig. 3　Learning curve in the 3 × 3 maze task. Since the curve is influenced by the random exploration, the performance change only by the exploration factor using the connection weights after learning is plotted for comparison.

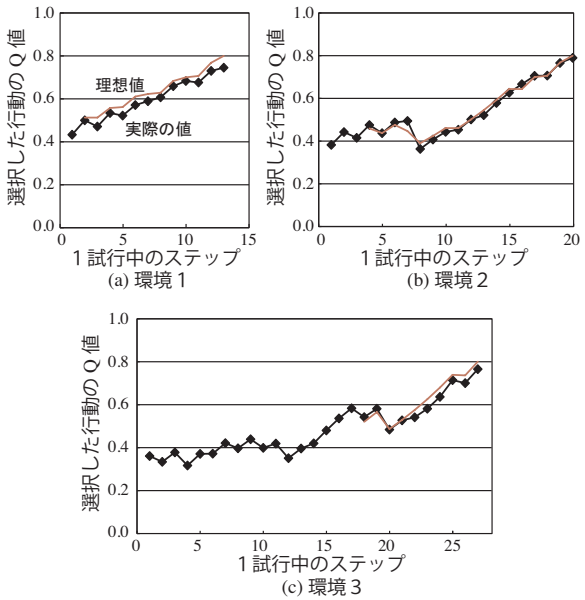

図 4　学習後の図 2 の 3 つの場合における試行中の最大 Q 値の変化．プロット点のない線は理想 Q 値を示す．

Fig. 4　Change of the maximum Q-value in one episode after learning for the 3 cases in Fig. 2. The line with no plot marker indicates the ideal Q-value.

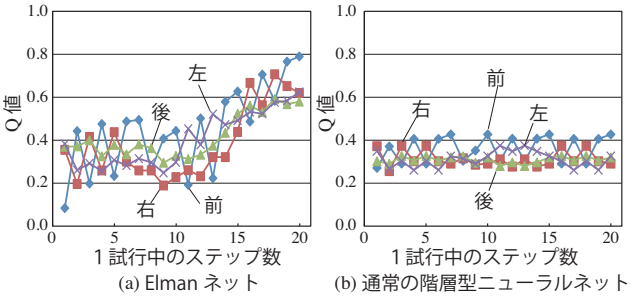

図 5　環境 2 の場合の，Elman ネットと通常の階層型ネットでの Q 値の変化の比較．

Fig. 5　Comparison of the Q-value change in the case of the environment 2 between using an Elman network and a non-recurrent feedforward network.

習後の実際の Q 値も単調に増加し，最後に報酬の値である 0.8 に近づいている．環境 2 では，エージェントは一方の行き止まりで反転する必要があるが，反転後 4 ステップはゴールの可能性がないマスを通過する．したがって Q 値は最初増加するが，8 ステップ目に急に減少し，その後再び上昇する．環境 3 では，エージェントは中央のマスから見て 4 つのすべての方向を探索するようになっている．方向を変えるために 1 ステップかかるため，エージェントは最初に中央のマスに戻ってきたときにはまっすぐ進み，次に戻ってきたときは方向を変え，3 回目に戻ってきたときには再びまっすぐ進み，25 ステップですべてのマスを訪問する経路が最適となる．しかし，学習後のエージェントは図 2(c) のような wall following の経路を取り，すべてのマスを訪れるのに 27 ステップかかった．

初期重み値と学習時の確率的探索のための乱数系列を変えて 10 回のシミュレーションを行ったが，常に最適経路を通ったものはなかった．上記の結果を含めて 10 回中 2 回は最大ステップが 28 となった．5 回は最大ステップが 40 ステップより小さく，残りの 1 回は 100 ステップより大きかったが，1,000 ステップよりは小さかった．リカレントでない通常の 3 層ニューラルネットを用いた場合，10 回のシミュレーションのすべてにおいて環境 3 の場合に無限ループに陥った．

図 5 は，環境 2 でのリカレントネットと通常の階層型ニューラルネットでのすべての Q 値の試行中の変化を比較したものである．どちらの場合も行動選択は適切であったが，リカレント構造がない場合は時間が進んでも Q 値の上昇は見られない．逆にリカレントネットの場合は，ゴール出現の可能性を考慮して Q 値が変化している．エージェントは周りの壁の配置と前の動作しか知ることができないにも関わらず，最後の 9 番目のマスに来ると，そこはゴールであり，報酬が得られることをわかっているようで面白い．また本来は，すべての壁配置に対して “wall following” にしたがって行動すれば記憶能力がなくても有限回でゴールに到達できるはずであるが，通常のニューラルネットの場合には，環境 3 では状態混同によって不適切な Q 値となり，適切な行動選択ができなかったものと考えられる．

3.3 未知環境での行動

次に，3×3 の迷路で学習したエージェントを学習中に経験していない環境に置き，獲得した知識がその環境に対してどれぐらい有効かを観察した．まず，9 個のマスが一列になっている環境にエージェントを置いた．一番左から 2 番目のマスにエージェントを置いた場合の行動を図 6 に，その際の選択した行動の Q 値のステップごとの変化を図 7 に示す．

エージェントは右端までまっすぐ進み，向きを変え，再び左端のまだ訪れていないマスのところまでまっすぐに進んでいることがわかる．Q 値は最初増加し，右端に着くと急に減少する．その後，Q 値は左端のマスに着くまで再び増加する．Q 値は最終的に 0.8 まで届かないものの，右端に着くとゴールはしばらく現れないことは反映されているように見える．環境 2 では，エージェントが右下のマスに行くとき，そのマスがゴールである確率が 20%であるのに対し，この場合は，右端のマスがゴールでなければ，ゴールの可能性は，訪れていない左端のマス 1

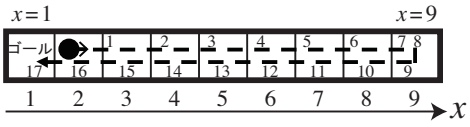

図 6　学習時に経験していない 9 マスを一列に並べた環境でのエージェントの行動．

Fig. 6　The agent behavior in an unexperienced environment where 9 squares are allocated in a row.

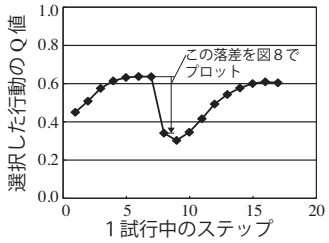

図 7　図 6 の環境の場合の試行中の最大 Q 値の変化．

Fig. 7　Change of the maximum Q-value in one episode after learning for the environment in Fig. 6.

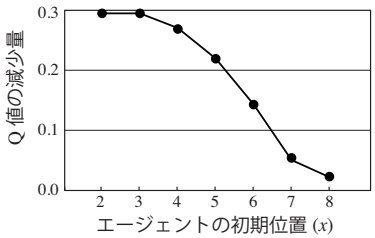

図 8　初期位置の違いによる，図 6 の右端での Q 値の減少量の差．

Fig. 8　Change of the Q-value decrease at the right end in Fig. 6 according to the initial agent location.

個残っているだけなので，ゴールである確率は 50%となる．したがって，図 4(b) の環境 2 での Q 値の減少よりも減少量が大きいことも合理的と思われる．また，スタート位置を右に移動させると，図 8 のように，右端に到着した際の Q 値の減少量（図 7 参照）は小さくなり，これも合理的と思われる．

次に，9 個のマスを図 9 のようにジグザグに配置した．また，エージェントが $(0, 1)$ のマスを右向きの状態でスタートしたときの行動も合わせて示した．試行中の最大 Q 値の変化を図 10 に示す．このグラフには，$(3, 4)$ のマスで下向きの状態でスタートした場合も一緒に示した．

エージェントの行動は最適な行動であった．右上端での Q 値の減少は小さいが，その前後で Q 値は増加している．また，右上端で向きを反転する前は，エージェントが前進して新しいマスに移動した場合はゴールである可能性があるが，右回転の行動を選んだ場合はゴールの可能性はないことが，図 10 で実際に Q 値が上がったり下がったりする変化と一致する．一方，向きを反転した後は，どんな行動をとってもすでに訪れたマスであるためしばらくは報酬を得られないことが，Q 値が最後のゴールに向かって単調に増加する変化と一致しており，エージェントは，向きを変えた後しばらくはゴールが現れないということを理解しているように見える．ただし，22 ステップ辺りから Q

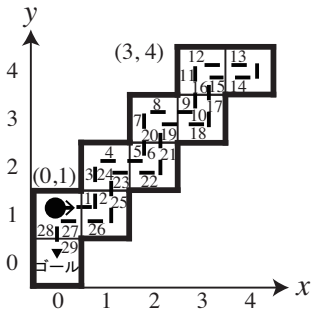

図 9　9 個のマスをジグザグに配置した学習時に未経験な状況でのエージェントの行動.

Fig. 9　The agent behavior in an unexperienced environment where 9 squares are allocated in a zigzag manner.

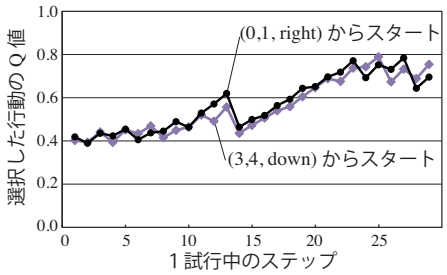

図 10　図 9 の環境での試行中の最大 Q 値の変化.

Fig. 10　Change of the maximum Q-value in one episode after learning for the environment in Fig. 9.

値の上下が見られるが，Q 値は早い段階で 0.8 になっており，未経験の環境で汎化がきかず，ゴール位置の判断が正しくできなかったためではないかと考えられる.

エージェントが $(3,4)$ のマスをスタートして向きを反転する前のニューラルネットへの入力は，$(0,1)$ のマスをスタートして向きを反転した後と同じである．同様に，$(0,1)$ をスタートして向きを反転する前の入力は，$(3,4)$ をスタートして向きを反転した後と同じである．にもかかわらず，どちらの場合も反転前は Q 値が上下し，反転後は単調増加する傾向にある.

3.4　分岐位置記憶タスク

記憶の機能の獲得をより詳しく調べるために，図 11 のように，1 列に並んだ 8 個のマスと 1 個の分岐マスよりなるより簡単な環境で学習をさせた．分岐マスの位置は図 11 のように 2 番目から 7 番目のマスの中から毎試行ランダムに選び，ゴールの位置も左端を除くすべてのマスからランダムに選んだ．エージェントは左端のマスを上を向いた状態でスタートする．この場合学習がより安定であったため，より正確な Q 値を実現し，記憶の獲得を明確に示すために学習係数を大きくしたが，それ以外のすべてのパラメータは前のタスクと同じとした.

図 11 はまた，分岐位置が $x = 2$ の場合のエージェントの行動も示している．一見，先に分岐マスを訪れることが最適行動であるように見えるが，分岐マスを訪れるためにはエージェントの向きを変えるためにたくさんの行動が必要になる．したがって，割引率が $\gamma = 0.92$ の場合，累積割引報酬の観点からはどこに分岐マスがあってもまっすぐに進み，分岐マスにゴールがある場合は右端で反転して戻る経路が最適となる.

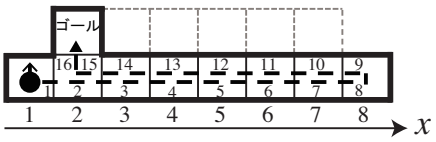

図 11　ランダム分岐のシミュレーション環境と学習後のエージェントの行動.

Fig. 11　Simulation environment with a random branch and a sample agent's behaviors after learning.

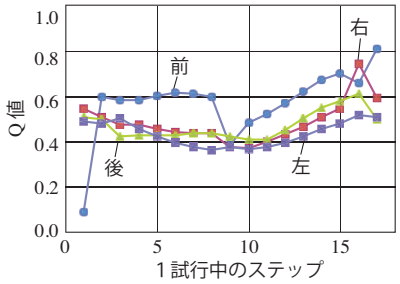

図 12　分岐タスクでの，試行中のすべての行動の Q 値の変化 (分岐位置 $x = 2$).

Fig. 12　Change of Q-values for all the actions in the branch task when the branch position is $x = 2$.

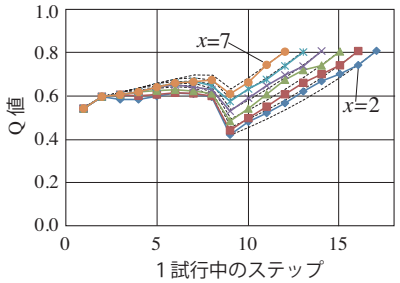

図 13　分岐位置による Q 値の変化の比較．破線は理想 Q 値.

Fig. 13　Comparison of the Q-value change due to the difference of the branch location. The broken line indicates ideal Q-value change.

図 12 は，分岐位置が $x = 2$ の場合の学習後のすべての Q 値の変化を示している．初期状態で右に回転した後，右端に着くまで前進の Q 値が最大となっている．右端に着いた 9 ステップ目で前進の Q 値はいったん減少するが，その後，分岐の位置に来るまで増加している．分岐のところでは右回転の Q 値が最大になり，最後の行動に対する Q 値はほぼ理想である 0.8 の値になっている.

図 13 は，6 個の分岐位置のそれぞれの場合について，試行中に選択した行動の Q 値の変化を示している．この変化は，プロットのない破線で表した理想 Q 値の変化と似ていた．エージェントが右端に着くまでゴールできない場合，9 ステップ目で Q 値は一旦下がっている．しかし興味深いことに，9 ステップ目での Q 値は，その時点での外部からの入力は同じであるが，分岐位置が折り返し位置から近いほど大きな値となっている．このことは，エージェントがリカレントネットの中に分岐位置を記憶し，それを Q 値に反映させるように学習したことを

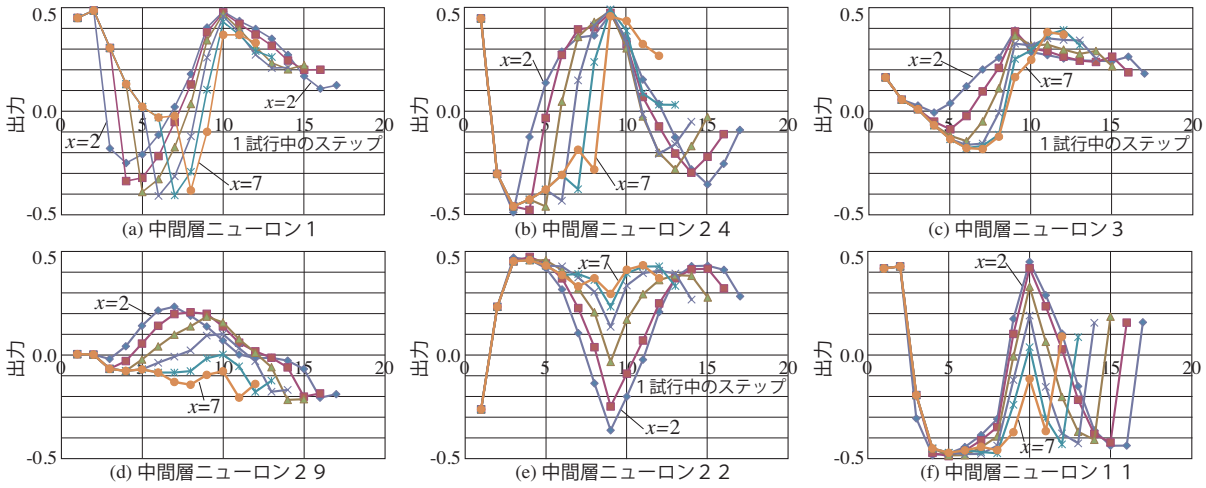

図 14　分岐位置の差によるいくつかの中間層ニューロンの出力変化の比較.

Fig. 14　Comparison of the output change due to the difference of the branch location in some hidden neurons.

表している．この場合，最適行動を実現するだけであれば記憶は必要ない．実際このタスクの場合，リカレントネットでなくても最適行動を学習することはできた．つまり，正確な Q 値の実現のために分岐位置を記憶することを学習したと言える．

40 個の中間層ニューロンを観察すると，分岐位置の記憶に貢献しているニューロンがいくつか見つかった．図 14 はそのいくつかの例を示す．中間層ニューロン 1 (a) と 24(b) は分岐位置に反応しているが，9 ステップ目で右端に到着した後は分岐位置によらず同じような値となっている．中間層ニューロン 3(c) や 29(d) はその値をしばらくの間保持しているように見える．中間層ニューロン 22(e) や 11(f) は，分岐位置での出力の差はあまり大きくないが，右端では分岐の位置を表現している．このように，強化学習を通して，必要に応じてニューロン間の値のリレーによって多値の記憶をするようになったことは，文献 [11] [12] と同様の傾向であった．またこの場合，最適な行動を学習するためには必ずしも分岐位置の記憶をする必要はなかったが，正確な Q 値を得るために，他から指示されることなく自律的に多値の分岐位置を記憶することを獲得した．

4. ま と め

見えないゴールを確率的な行動によらないで探索する 3 × 3 のランダム迷路タスクにおいて，エージェントはリカレントネットを用いた強化学習を通して，最適ではないものの適切な行動を獲得した．学習後の Q 値は理想的な値に近かったが，通常の階層型ニューラルネットではそれを実現できなかった．学習後のエージェントは，行き止まりのところで向きを変えるとしばらくはゴールがないことを理解しているように見えた．さらに，報酬だけからの学習を通して，より正確な Q 値を実現するために，分岐位置という多値情報を記憶する能力がリカレントネットの中に創発することを確認した．

Acknowledgment

本研究は，科学技術研究費補助金#19300070，#23500245 の補助を受けた．ここに謝意を表する．

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, Vol. 48, No. 1, pp. 106–111, 2009

[2] K. Shibata, Emergence of Intelligence through Reinforcement Learning with a Neural Network, Advances in Reinforcement Learning, A. Mellouk (Ed.), InTech, pp. 99–120, 2011

[3] S. Thrun, Efficient Exploration In Reinforcement Learning, Tech. Report CMU-CS-92-102, Computer Science Department, Carnegie Mellon University, 1992

[4] G. Zhao, S. Tatsumi & R. Sun, RTP-Q: A Reinforcement Learning System with an Active Exploration Planning Structure for Enhancing the Convergence Rate, Proc. of IEEE SMC'99, fp063.pdf, pp. V-475-480, 1999

[5] K. Shibata, Acquisition of Deterministic Exploration Behavior by Reinforcement Learning, Proc. of the 11th Int'l Symp. on Artificial Life and Robotics (AROB), 2006

[6] K. Shibata, Learning of Deterministic Exploration and Temporal Abstraction in Reinforcement Learning, Proc. of SICE-ICCAS, pp.4569-4574, 2006

[7] 後藤健太, 柴田克成, リカレントネットを用いた強化学習による記憶を利用した探索行動の学習, 第 61 回電気関係学会九州支部連合大会講演論文集, CD-ROM, 2008

[8] K. Goto & K. Shibata, Acquisition of Deterministic Exploration and Purposive Memory through Reinforcement..., Proc. of SICE Annual Conf. 2010, FB03-1.pdf, 2010

[9] B. Bakker, et al., A Robot that Reinforcement-Learns to Identify and Memorize Important Previous Observations Proc. of IROS 2003, 430-435, 2003

[10] H. Utsunomiya & K. Shibata, "Contextual Behavior and Internal Representations Acquired ...", Adv. in Neuro-Information Processing, LNCS, Vol. 5507, pp. 970-978, 2009

[11] 後藤健太, 松本康生, 柴田克成, リカレントニューラルネットを用いた強化学習による予測機能の創発, SICE 九州支部学術講演会, pp. 81-84, 2009

[12] K. Goto & K. Shibata, Emergence of prediction by reinforcement learning using a recurrent neural network, Journal of Robotics, Vol. 2010, Article ID437654, 2010

[13] D.E. Rumelhart, G.E. Hinton & R.J. Williams, Learning Internal Representations by Error Propagation, Parallel Distributed Processing, The MIT Press, pp. 318-362, 1986

[14] C.J.C.H.Watkins & P. Dayan, Technical Note:Q-Learning, Machine Learning, Vol.8, pp.279-292, 1992

[15] R. S. Sutton & A. G. Barto, Reinforcement Learning, MIT Press, 1998