強化学習により対象物検出行動を学習した画像入力ニューラルネットにおける中間層ニューロンの解析

Analysis of Hidden Neurons in the Neural Network with Image Input that Learned Target Detecting Actions by Reinforcement Learning

大分大学　○柴田克成，河野友彦

○ Katsunari Shibata and Tomohiko Kawano　Oita University

Abstract: In the previous work[1], the authors showed that by reinforcement learning with a neural network whose inputs are 6240 color image signals, a robot could learn to turn its camera to a target and bark with about 90% success ratio in a real-world-like environment with various backgrounds and light conditions. In this paper, the neural network after this learning was analyzed. Many hidden neurons seemed to work as a template to detect specific target positions, and division of roles among hidden neurons were formed. It was also confirmed that the internal representation to indicate the presence of target that is uninfluenced by backgrounds and light conditions was acquired through learning. These results suggest that the simple and general learning system has an ability to obtain purposive internal representation autonomously through learning, and has a potential for avoiding the frame problem[2].

1　はじめに

「知能ロボット」研究は長年行なわれてきているが，こと高次機能に関しては，あまり研究が進んでいないというのが現状であろう。われわれは，「たくさんの知識を与えられたロボットが本当に賢いと言えるのか？」との疑問に端を発し，いかに少ない知識で諸機能を学習させられるかを追究してきた。また，ロボットがフレーム問題 [2] を回避し，人間のように柔軟に学習を行うためには，従来のように，機能モジュールに分割し，それをシーケンシャルに接続するという開発方法からの大きな方向転換が求められる。そのためには，生物の脳のように超並列なシステムを用意し，全体が調和を保ちつつ柔軟に学習できること，さらに，ロボットの目的は，センサからモータまでの処理を何らかの目的の下で最適化することであり，人間がそれを理解することではないという観点から，できるだけ設計者の干渉を排除し，システム自身の最適化に任せることが重要であると筆者らは考えている。

このような考えに基づいて，筆者らは，Fig. 1 のように，ロボットのセンサからモータまでを１つのニューラルネットで直接つなぎ，強化学習でそれを学習させるという，一見効率の悪い手法により，ニューラルネット内に必要となる認識，記憶，会話，社会行動などの諸機能を，合目的的，適応的かつ調和的に創発させることを提唱し，簡単なタスクのシミュレーションを通してその可能性を示してきた[3][4]。高次機能は，認識や制御といったセンサやモータに近い部分と違い，何を入力とし，何を出力とするかを予め設定することが困難である。これに対し，提案手法によって，予め入出力を与えることなく柔軟な高次機能の形成が実現できれば，高次機能研究における閉塞状態の打破につながるのではないかとの期待も込めて研究を進めてきた。

しかしながら，強化学習やニューラルネットの個々の学習能力はある程度認められても，センサとモータ間をニューラルネットでつなぎ，何の知識の付与もせず強化学習で学習するだけでは，とても実世界での複雑なタスクの学習は

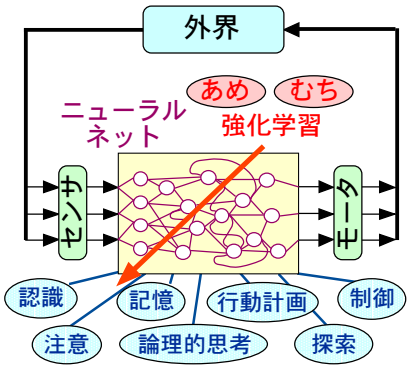

Fig. 1: Parallel and cohesively flexible learning of whole process from sensors to motors by the coupling of RL and NN.

不可能であるとの見方が根強い。これに対し，実ロボットを用いた実験による検証も進めてきた。そして，1000 個以上のモノクロ視覚信号をニューラルネットへ直接入力し，黒い箱にアプローチして押すという行動を学習によって獲得できることを示した [5]。また，より実世界に近い環境での有効性を示すため，周りを囲んでいた白い紙を取り払い，さまざまな照明条件，さまざまな背景の中で，6000 個を越えるカラー画像信号をニューラルネットへ直接入力し，可動カメラとしてのロボットが対象物を正面で捉え，それを正しく認識できた場合に報酬を与えるという設定で学習させ，学習に用いていない画像を提示しても，90%程度の成功率で正しく認識できることを示した [1]。

本論文では，この学習を行なった後のニューラルネットの中間層に形成された内部表現について解析し，強化学習とニューラルネットの組み合わせによる合目的的かつ自律的な学習能力を示す興味深い結果が得られたので報告する。

2　強化学習とニューラルネットの組み合わせ

強化学習とニューラルネットを組み合わせる手法は非常に簡単である。センサから得られる信号を直接ニューラル

ネットに入力し，計算された出力に対し，強化学習に基づいて生成された教師信号でニューラルネットを教師あり学習させる。過去のセンサ信号を考慮する必要がある場合は，ニューラルネットとしてリカレント型のニューラルネットを用いることもできるが，ここでは，通常の階層型のニューラルネットを用いる。また，強化学習としては，actor-critic[6] や Q-learning[7] を用いることができるが，離散行動を対象とする本研究では，それに適した Q-learning を用いたので，ここではその場合の方法について説明する。

ニューラルネットの出力層には，ロボット等の行動の数だけのニューロンを用意し，その出力をそれぞれの行動に対する Q 値として扱う。そして，現在のセンサ信号をニューラルネットに入力して，出てきた Q 値としての出力を見て，行動を決定する。行動決定の方法は，通常の Q-learning と同様な方法を取ることができる。本稿では，ϵ-greedy[8] を用いる。また，ここでは，エピソディックなタスクを想定し，エピソード終了時のみ報酬や罰が与えられるものとする。エピソード終了時には，そのときの行動 a_t に対応した Q 値の出力に対する教師信号を

$$T_{a_t,t} = r_{t+1} \tag{1}$$

とする。ただし，r_t は強化信号 (報酬や罰) である。また，通常時には，選択された行動を実行した後の時刻 $t+1$ に得られたセンサ信号 S_{t+1} をニューラルネットに入力し，そのときの最大の出力を用いて，一つ前の時刻 t の行動 a_t に該当する Q 値の出力に対する教師信号 $T_{a_t,t}$ を，

$$T_{a_t,t} = \gamma \max_a O_a(S_{t+1}) \tag{2}$$

と計算する。ただし，γ は割引率，$O_a(S)$ はセンサ信号 S を入力した際の行動 a に対応するニューラルネットの出力を示す。そして，時刻 t のセンサ信号 S_t を再度入力して出力計算を行なった後に，この教師信号でニューラルネットを通常の誤差逆伝播 (BP) 法で学習させる。該当しない行動の出力は，学習させなかった。また，実際には，ニューラルネットの値域と Q 値として表現したい値の範囲がずれる場合があるので，両者の間を適宜線形変換する。ここでは，各ニューロンの出力関数であるシグモイド関数の値域を-0.5 から 0.5 としたが，出力-0.5 を Q 値の-0.1，出力 0.5 を Q 値の 0.9 とするため，出力から Q 値へは 0.4 を足し，逆の場合には 0.4 を引いた。

3 実験概要 [1]

3.1 タスク設定と学習方法

実験では，2 台の SONY 製犬型ロボット AIBO を用いたタスクを行った。両者は Fig. 2 のようにステーションに置いて 43cm 離して向かい合わせに固定し，そのうち 1 台 (黒) は首を振ることで可動カメラとして用い，もう 1 台 (白) は少し複雑な形をした動かない対象物として用いた。カメラは水平画角は約 54° であり，原画像として 208×160 の RGB データが得られるが，AIBO から送信する前に XY 方向で 4 ピクセルごとに間引き処理を行い 52 × 40 と 16 分の 1 のサイズに縮小したものを学習に用いた。カメラの水平可動範囲は，正面を向いている時を 0° として ±89.6° であるが，カメラ画像の範囲内に対象物が収まるように，±20° の範囲を 5° の間隔でスイングさせた。そのため，首の位置の状態数は 9 個となる。しかし，実ロボットを用いているため，対象物の位置は少しずれることがあった。

黒の AIBO は，画像をニューラルネットに入力した際の出力に従って行動する。行動は，“ 首を右に動かす ” 首を左に動かす ” 吠える ”の 3 行動とし，首を動かす場合には該当の方向へ 5° 回転させた。行動選択における ε -greedy での探索率 ε は 0.13 で固定とした。初期の首の位置は，1 試行ごとに 9 状態の中からランダムに決定した。そして，首を動かし，もう 1 台の AIBO が正面に見えている状態 0 のときに吠えると報酬がもらえ，それ以外のときに吠えると罰を受ける。たとえば，Fig. 3 のように正面より右に 10° の方向を向いている状態 2 から開始する場合では，首を左に 2 回動かして，正面を向いてから吠えるという行動をとると報酬をもらうことができるが，状態 1 や状態 2 で吠えてしまうと罰を受けることになる。報酬の値は 0.8 とし，罰は，現在の Q 値から 0.02 を引いたものとした。

ニューラルネットへの入力は，52 × 40 の 2080 画素に，RGB の 3 を掛けた計 6240 個の信号とした。RGB それぞれの 256 階調を 0.0 から 1.0 に正規化した後に，0, 1 を反転した値を入力した。ニューラルネットの出力は，行動の数と同じ 3 つ用意した。ニューラルネットは階層型で 5 層とし，それぞれのユニット数は，最下層から 6240-600-150-40-3(約 384 万結合) とした。各ニューロンの出力関数は，値域が-0.5 から 0.5 のシグモイド関数を使用し，学習係数は 0.5 とした。式 (2) における割引率 γ は 0.8 とした。

実ロボットを使った学習は非常に時間がかかるため，学習用にさまざまな画像をあらかじめ採取しておき，実機を用いずに学習を行った。画像の採取時は，外が明るい昼間，夕方，暗くなった夜，ブラインドの開閉などの照明条件を変え，さらにロボットの後ろにさまざまな物や壁紙を適当

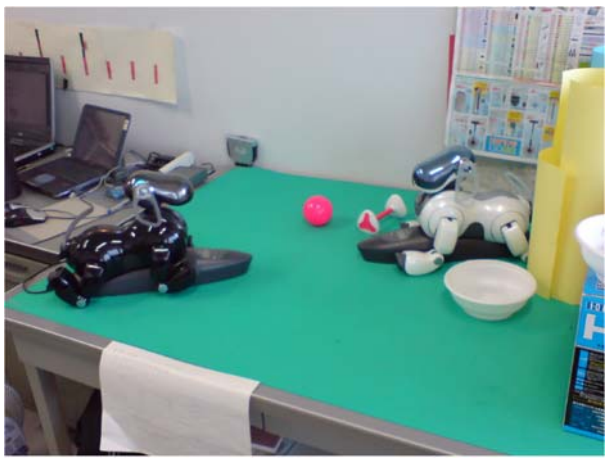

Fig. 2: The environment for the experiment.

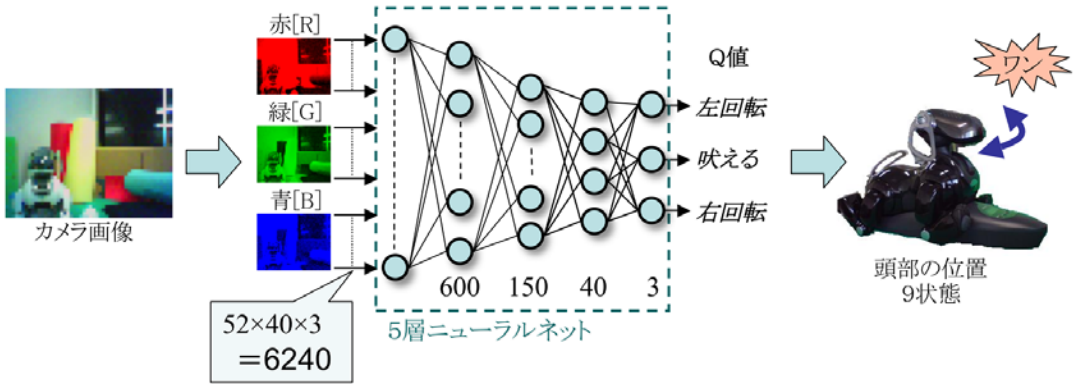

Fig. 4: The processing in the black AIBO.

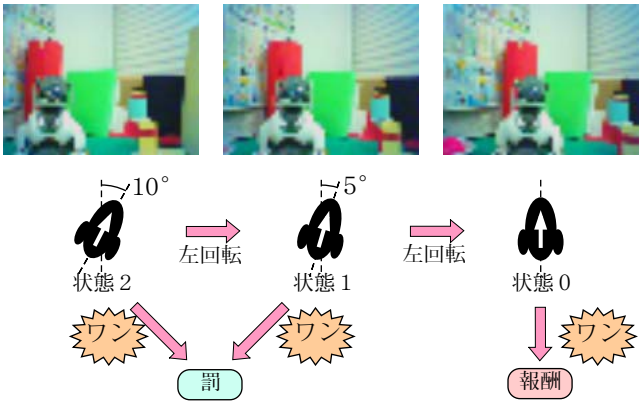

Fig. 3: Sample of the state transition in the experiment.

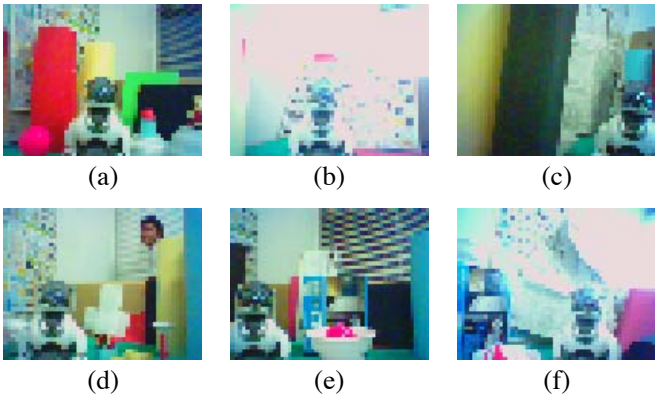

Fig. 5: Some sample images captured under different light conditions and backgrounds that were used in learning.

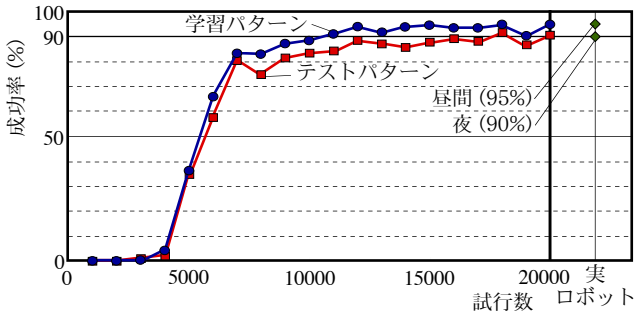

Fig. 6: The learning curve of success ratio.

に配置した。そして，首の 9 個の状態それぞれに 1 枚ずつの画像を採取して 1 パターンとし，条件を変えることで，同じ状態でもパターンごとに明るさ，背景が異なるようにした。学習用サンプルとして 312 パターン 2808 サンプル，パフォーマンスのテスト用に 92 パターン 828 サンプルを用いた。学習時は，ステップごとに 312 パターンの該当状態の画像の中からランダムに選択した。Fig. 5 に，実際に学習に用いた画像の例を示す。背景，照明条件によって画像が大きく変化することがわかる。100 試行の学習ごとに，$\varepsilon = 0$ (ランダム成分を 0) で学習しない状態で 50 試行のパフォーマンステストを行い，1000 試行ごとにその平均をとった。このテストでは，学習に使用したものと同じ訓練パターンと，汎化能力の指標とするために，学習に全く使用していないテストパターンで別々にテストを行った。テストでの初期位置は，学習時と異なり，試行ごとにゴールから一番遠い状態 -4 と状態 4 に交互に切り替えた。

3.2 実験結果

ここでは簡単に実験結果を述べる。詳細は，文献 [1] を参照されたい。テスト時は，ゴールから一番遠い状態からスタートしているため，Fig. 6 のように，学習開始当初は，ほとんど正しく吠えることはなかったが，4000 試行あたりから学習成功率が上昇し始め，10000 試行を越えたあたりで学習パターンでの成功率は 90%を越えた。20000 試行の学習後，学習パターンでは 95%程度，未知のパターンであるテストパターンの場合でも 9 割程度まで成功率は上昇しており，ニューラルネットの汎化能力が非常に有効に働いていることを確認した。シミュレータによる学習時に得られた重み値を使用し，100 試行だけ実機で動かしてみた。シミュレータでは 1 ステップごとにパターンをランダムに変化させていたが，実機では 20 試行ごとに背景を変えた。つまり，100 試行行うので 5 パターンの背景を与えたことになる。また実験中は，20 試行ごとにブラインドの開け閉めも行い照明条件を変えながら行った。また，外がまだ明るい昼間と真っ暗になってからの夜とで 2 回実験を行った。その結果，昼間でも夜でも 9 割程度の検出率となっており，シミュレータでの学習によって実機でもある程度検出行動を行うことができることを確認した。

4　中間層ニューロンの解析

4.1　個々の中間層ニューロンの働き

次に，一番下の中間層の 600 個のニューロンのそれぞれについて，入力からの結合の重み値を観察した。各ニューロンへの結合の数は，入力ニューロンの数，つまり，入力画像の画素数 × RGB である 6240 と一致するため，各ニューロンへの重み値の絶対値の最大値を用いて重み値を 0 から 255 に線形変換し，該当画素の諧調とすることで，各ニューロンへの重み値の分布を 1 つの画像として表現した。しかし，その画像は，Fig. 7(c-3) のように，ただのランダムなパターンにしか見えなかった。そこで，重み値そのものではなく，学習前と後での重み値 w_{before}, w_{after} の変化量を

$$pixel_{color,i,j} = (int)\left(\frac{w_{after,color,i,j} - w_{before,color,i,j}}{\max_{color,i,j}\left|w_{after,color,i,j} - w_{before,i,j}\right|} \times 127\right) + 128 \quad (3)$$

の式にしたがって 0 から 255 に線形変換して画像化した。ただし，*color* は該当画素の色であり，R, G, B のどれかが入り，*i*, *j* は該当画素の行と列の番号を表す。重み値の値が学習によって増加すると，その重み値に対応する画素の，対応する色は明るくなる。Fig. 7(c-3) に示したニューロンの場合，重み値の絶対値の最大は，学習前が 0.127，学習後が 0.131 であった。他のニューロンもほぼ同様な値であった。一方，学習による重み値の変化量の絶対値の最大は 0.012 と小さかったが，初期重み値を除いたそのわずかな変化量を，式 (3) を通して画像として表現したものが Fig. 7(b-3) となる。このように，学習後の中間層の表現に初期重み値の影響が大きく残ることは，文献 [10] でも示されている。

作成された各最下層の中間層ニューロンに対応した 600 個の画像を見ると，多くの場合，画像中に AIBO の姿を見つけられた。その例を Fig. 7 (b-1,..,6) に示す。Fig. 7 (a-1,2,3) には，AIBO が映った実際の画像の例を示す。Fig.(b-1,2,3) に現れる AIBO の位置は，実際の画像の Fig.(a-1,2,3) に表れる AIBO の位置とほぼ同じ位置であることがわかる. これは，最下層の中間層ニューロンが，特定の位置に AIBO がいるかどうかをみつけるためのテンプレートとして働いていると解釈することができる。それぞれのニューロンごとに表現が異なっており，強化学習を通して自律的にニューロン間で役割分担ができていることが大変興味深い。また，Fig. 7 (b-4) では，左半分に AIBO のネガのパターンが，真ん中あたりにポジの AIBO のパターンが見える。Fig. 7 (b-5) では，少しぼやけたネガの AIBO が真ん中あたりに，両端にポジの AIBO が見える。Fig. 7 (b-6) では，１つの AIBO の左半分がネガ，右半分がポジに見える。これらの中間層ニューロンは，照明条件に影響されない AIBO の認識に役立っているのではないかと推測することができる。Fig. 7 (b-7) は，真ん中の中間層のニューロンの一つについて，最下層の中間層ニューロンへの重み値の変化を，最下層と真ん中の中間層ニューロン間の重み値によって重み付けした後に，0 から 255 に正規化したものである。ここには，少し太った AIBO の姿が見える。これは，AIBO の首の制御があまり正確にできないために生じる画像のずれを吸収しているのではないかと推測される。また，これらはあくまでも推測である。実際の脳の機能が簡単に理解できないように，実際には並列性を利用した，われわれには理解が困難な巧みな処理を行なっていると考えられる。

4.2　強化学習による内部表現の変化

最後に，ニューラルネットが背景や照明条件によらない AIBO の位置を表現する内部表現を持っていることを示すためのシミュレーションを行った。まず，学習後のニューラルネットの出力ニューロンを取り除き，新しく 1 つの出力ニューロンを設け，最上層の中間層ニューロンと重み値 0 で結合させる。そして，２つの AIBO の相対位置をずらさないように，学習時とは別の場所に移し，Fig. 8 の (a)(b) それぞれにある 12 個の新しい画像を取り込んだ。これらは，２つの画像 (上段と下段) で１つのペアをなし，同じ背景，照明条件で，正面に AIBO がいるかいないかだけが違うものとした。3 つのペアでは，背景には何も置かず，その他の３つのペアでは，背景にいくつかの物体を置いた。また，３つのペアは，それぞれ，ブラインドを開けた状態の昼間，ブラインドを下ろした状態の昼間，夜に画像を取得した。そして，この 12 個の画像のうち，Fig. 8 で，“learned ’’というラベルがつけられた２つの画像のみを訓練パターンとし，その他の 10 パターンをテストパターンとした。2 つの訓練パターンのうちの 1 つでは，AIBO の後ろには何も置かれておらず，昼間取った画像であるため画像が明るい。一方，もう一つの方は，AIBO はおらず，背景にいくつかの物体が置かれ，かつ，夜取得した画像のため，画像は全体的に暗い。2 つのうちの 1 つが毎回ランダムに選択され，前者に対しては 0.4 を，後者に対しては-0.4 を教師信号として与えて教師あり学習させた。

学習後，学習に用いた 2 つの画像に対する出力は，それぞれ教師信号とほとんど同じ値となった。そして，それ以外の 10 個の画像を入力したときの出力の値を観察した。そして，この教師あり学習の前に強化学習を行った場合と行わなかった場合で出力の分布を比較した結果を Fig. 8 に示す。(a) が強化学習を行っていない場合，(b) が強化学習を行った後の場合である。両者で，すべての初期重み値は同じとした。強化学習を行っていない方では，出力が背景や照明条件によって大きく異なっていることがわかる。画像の信号としては，画像中に AIBO が占める面積は小さいため，AIBO が存在するかしないかよりも，背景や照明条件の方に大きく影響される。したがって，ニューラルネットの汎化能力という点から考えると，AIBO の存在の有無よりも背景や照明条件によって出力がより大きく変化することは自然な結果である。一方，強化学習後のニューラルネットでは，出力の符号が AIBO の存在の有無によって決まっていることがわかる。このことは，強化学習を通してニューラルネットが，背景や照明条件にかかわらず画像の中心に AIBO がいるかどうかを表現できるようになったこ

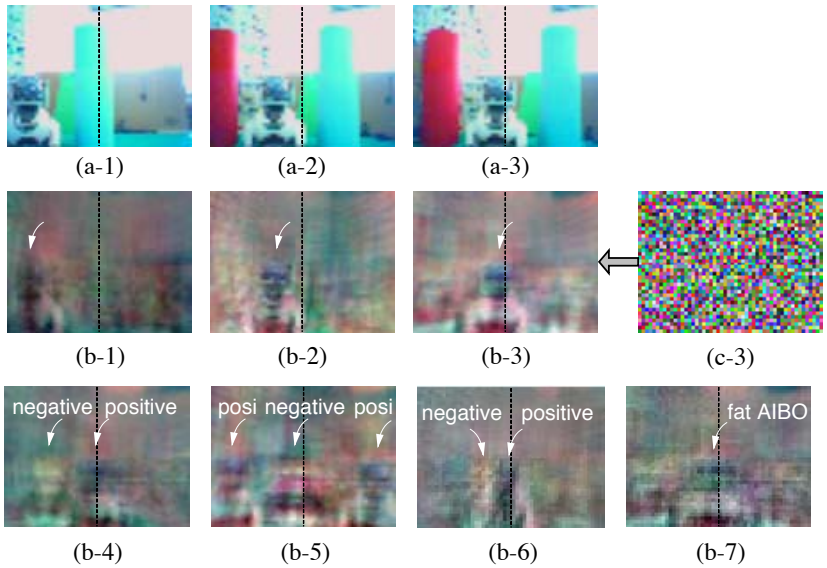

Fig. 7: (a) Samples of actual image, (b) Visualization of the change of the connection weights from the input layer for 5 lowest hidden neurons. AIBO image can be found at the tip of each small white arrow. Only (b-7) is the image of one neuron in the middle hidden layer.

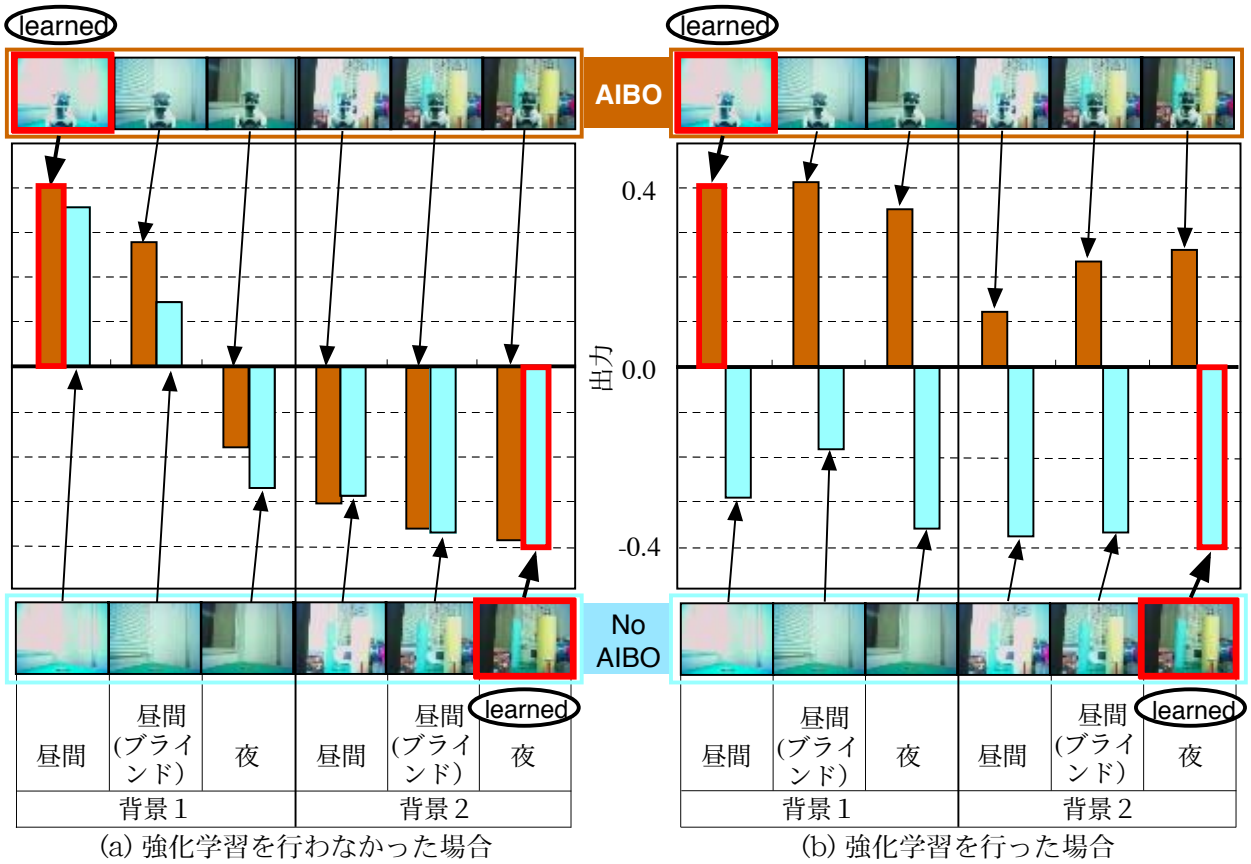

Fig. 8: One output neuron was added to the neural network(NN) after reinforcement learning(RL), and supervised learning was performed using only two input images labeled as “learned”. The training signal for one pattern is 0.4, while it is -0.4 for the other. After that, the output for 10 test input images was observed. The light condition, background, and also the existence of AIBO are varied in the input images. The outputs are compared with (a) the case of using the NN before RL and (b) the case of using the NN after RL. In the latter case, the output changes according to the existence of AIBO.

とを意味する。そして，汎化能力が，入力であるセンサ信号空間の上ではなく，獲得した中間層ニューロンの表現の上で働いていると考えることができる [10]。

この実験では，黒い AIBO に対して，誰も白い AIBO の場所を認識しなさいと教えていないし，どうやって背景や照明条件によらずに認識するのかも教えていない。にもかかわらず，学習を通して，必要に応じてそのような機能を獲得した。

5 議論

このタスクは，6240 個の信号からなる 2808 個の画像を 3 つに分類する問題として捉えると，一見簡単そうに見える。しかし，われわれはこのタスクの目的を予め把握しており，首の位置が 9 個に限定されていることも予め知っている。また，われわれはテンプレートマッチングという手法を知っており，かつ，明るさを補償すればこのような問題に有効に働くであろうと推測することができる。このようなことを最初から知っているからこそ，この問題を簡単だと考えることができるわけである。一方，AIBO はそのようなことを全く知らない状態からスタートしている。このことを考慮することは，自律的な学習能力を考える際には非常に重要である。また，分類問題と言っても，学習中に，吠えたときの報酬と罰以外に外部から与えられる情報はない。実際にどの画像の時にどの行動を取れば良いかのサンプルは与えられず，どう分類するか自身も学習を通して獲得しなければならない。さらに，ここで用いている学習システムの構造や学習方法は，この問題に特化したものではなく，非常に簡単で汎用的なものである。にもかかわらず，テンプレートのような表現を自律的に獲得し，背景や照明条件に影響されないような内部表現を獲得することができたということは，このシステムの自律性，柔軟性が非常に優れていることを表している。また，この問題を分類問題として，SVM(Support Vector Machine) を用いて学習できたとしても，SVM では，柔軟にその内部表現を形成することはできず，高次機能への拡張性が感じられない。

柔軟な知能ロボットの開発においては，フレーム問題が大きな問題となることが指摘されている [2]。これに対し Brooks は，Subsumption Architecture[11] を提案し，処理をシーケンシャルにつながずに，並列に配置することの有効性を示している。しかしながら，さまざまなタスクに対応できる柔軟なシステムにするために，そもそもどのような処理をどのように配置すれば良いかを決めることは，Brooks 自身も指摘しているように非常に難しい問題である。またこの場合も，反射的な行動は得意とするが，高次機能への展開の道筋はなかなか見えてこない。これに対し，本手法は，タスクに必要な要素技術をあらかじめ用意しておく必要はなく，自律的に，内部に機能が創発するため，このフレーム問題を本質的に解決する手段となり，また，入出力が事前に定まらない高次機能の創発にも力を発揮する可能性を感じさせるものであると考えている。

6 結論

センサからモータまでをニューラルネットでつなぎ，それを強化学習で学習させるという非常に簡単で汎用的なシステムによって，実世界に近い環境においても，テンプレートと考えられるような内部表現や背景や照明条件によらない対象物に関する情報の表現といったタスクに応じた柔軟な認識を内部に自律的に形成できることを示した。このような汎用的なシステムによる柔軟な機能創発は，フレーム問題の本質的な解決とともに，入出力を事前に決定することが困難な高次機能の学習による獲得へ向けた可能性を感じさせるものであると筆者らは考えている。

謝辞

本研究は，日本学術振興会科学技術研究費補助金 #193000701 の補助の下に行われた。ここに謝意を表する。

参考文献

[1] 河野友彦, 幸和芳, 柴田克成: 画像入力ニューラルネットを用いた強化学習による可動カメラの対象物検出行動, 第 26 回計測自動制御学会九州支部学術講演会予稿集, pp. 155-158, 2007

[2] D. Dennett, Cognitive Wheels: The Frame Problem of AI, The Philosophy of Artificial Intelligence, Margaret A. Boden, Oxford University Press, pp. 147–170, 1984

[3] 柴田克成, 岡部洋一, 伊藤宏司: ニューラルネットワークを用いた Direct-Vision-Based 強化学習, 計測自動制御学会論文集, Vol.37, No.2, pp.168-177, 2001

[4] 柴田克成, 岡部洋一, 伊藤宏司: ニューラルネットを用いた強化学習とロボットの知能,「ニューラルネットワーク計算知能」, 渡辺桂吾編著, 森北出版, 第 12 章, 2006

[5] 柴田克成, 飯田大: 視覚センサ付き実ロボットによる箱押し行動の獲得, 第 14 回インテリジェント・システム・シンポジウム (FAN) 講演論文集, pp. 123-128, 2004

[6] A.G. Barto, R.S. Sutton, and W. Anderson: Neuronlike adaptive elements can solve difficult learning control problems, IEEE Trans. on Systems, Man, and Cybernetics, Vol. 13. No. 5., pp. 834–846, 1983

[7] C.J.C.H. Watkins: Learning from Delayed Rewards, PhD thesis, Cambridge University, Cambridge, England, 1989

[8] R. S. Sutton and A. Barto: Reinforcement Learning: An Introduction, A Bradford Book, The MIT Press, 1998

[9] D. E. Rumelhart et al.: Williams: Learning Internal Representation by Error Propagation, Parallel Distributed Processing, MIT Press, Vol. 1, pp. 318–364, 1986

[10] 柴田克成, 伊藤宏司: 階層型ニューラルネットにおける中間層での適応的空間再構成と中間層レベルの汎化に基づく知識の継承, 計測自動制御学会論文集, Vol. 43, No. 1, pp. 54–63, 2007

[11] R. A. Brooks: Intelligence Without Representation. Artificial Intelligence, Vol. 47. pp. 139–159, 1991