解説：特集
高次機能の学習と創発
—脳・ロボット・人間研究における新たな展開—

強化学習とニューラルネットによる知能創発

柴 田 克 成*

*大分大学　大分県大分市大字旦野原 700
*Oita University, 700 Dannoharu, Oita, Japan
*E-mail: shibata@cc.oita-u.ac.jp(shibata@oita-u.ac.jp に変更予定)

キーワード：強化学習 (reinforcement learning)，ニューラルネット (neural network)，知能創発 (emergence of intelligence)，並列アーキテクチャ (parallel architecture)，無意識処理 (unconscious process).
JL 0001/09/4801–0106 ©2009 SICE

1. はじめに

「人間が画像認識，行動計画，制御などのプログラムを書いて与え，それを忠実に実行しているロボットを知能ロボットと呼んで良いのだろうか？」本稿では，この疑問を長年持ち続けてきた筆者が，「強化学習」と「ニューラルネット」の組み合わせが，思うように進まない高次機能研究に大きな風穴を開ける可能性を解説するとともに，その可能性の判断材料として，現在それで何ができるかを紹介する．

2. 高次機能研究に何が必要か？

ここでは，なぜ高次機能の研究が今まで思ったように進まなかったかを考察し，筆者が考える高次機能研究のあるべき姿を浮かび上がらせていく[1]．

従来の脳研究，知能ロボット研究を見ると，「機能モジュール化」という大きな共通点に気がつく．脳研究では，各領野，部位の働きを特定することが目的となっており，知能ロボット研究では，ロボットの処理を認識，制御などの機能モジュールに分け，個々の高機能化による全体の高機能化を目指している．その中で「高次機能」は，センサに近い認識部，アクチュエータに近い制御部とは違い，入力も出力も与えられない．さらに，人間のように，さまざまな言語の習得や新しい概念の形成などあらゆる機能を柔軟に獲得することが求められる．しかし，入出力を与えれば，柔軟性が失われることになり，結局どうすればよいかわからなくなる．これが「高次機能」研究がなかなか進まない大きな要因であり，その根源は，「機能モジュール化」にあると筆者は考えている．つまり，高次機能を実現するモジュールを他から分離して考えるからその入出力をあらかじめ決めなければならず，柔軟性も大きく奪われることになるのである．

では，われわれはなぜ「機能モジュール化」してしまうのであろうか？その根本的な要因は，われわれの「脳」と「意識」のずれにあると筆者は考えている．「脳」は何百億，1千億というニューロンよりなる超並列システムと言われ，かつ，非常に柔軟に変化するシステムでもある．一方，「脳」が生み出すわれわれの「意識」はシーケンシャルであり，かつ，言語的な表現をするため，「脳」のように並列で，柔軟に変化していくものを表現し，理解することが困難である．そこで，機能モジュールという枠で捉えることにより，シーケンシャルに流れを追えるレベルまで情報量を落とし，それを言語的に表現して理解に努め，ロボットの処理を構成する際は，それをシーケンシャルに配置することになる．

そう考えると，われわれが超並列で柔軟な「脳」を完全に理解することは，「意識」を通す限り不可能ではないかと考えられる．しかし，われわれは「意識」を通ってきたものしか理解できないため，脳の中で無意識に行っている超並列処理には気付かず，つい，「意識」で理解されたことが，脳が行っていることのすべてだと思いがちである．そして，脳が「行っている」のだから「理解できる」はずと思い込み，機能モジュール化して理解していけば，いつかは脳が分かり，モジュールの機能向上を図っていけば，いつかは人間のようなロボットができると期待してしまう．

無意識に行われている処理の存在は，視覚野での方位選択性セルの応答が意識に上らないことからも明らかであるし，錯視や choice blindness[2] と呼ばれる現象も，「脳」と「意識」のずれの結果と言えるだろう．ニューロン間の結合関係をすべて調べ上げれば，脳の「復元」はできるかもしれない．しかし，柔軟性まで含めた「復元」は困難だろうし，「理解」できなければ，「応用」は利かないだろう．

以上のことから，まず，**脳を完全に理解し，それに基づいて人間のような知能ロボットを開発することは不可能であることに気付き，現在の「機能モジュール化」のアプローチから大きな方向転換を図ること**が，高次機能研究には不可欠である．これが，本稿の最初のポイントである．さらに，無意識の処理まで実現させるためには，**超並列で非常に柔軟で，かつ，全体としての調和を保つことができるシステムの導入が必要である**．これが第2のポイントである．

Brooks の Subsumption Architecture[3] は，従来のシーケンシャルなシステムに対し，並列アーキテクチャの導入を提唱したもので，その機敏で柔軟な動きによって，フレーム問題[4] の回避に一定の役割を果たした．しかしながら，彼は複雑なシステムを分解して考えること，そして，人間がそのシステムを理解できることが重要であると述べ，レイヤーと呼ぶ機能モジュールを並列に配置し，そのプログラムの開発と，レイヤー間のやりとりの設計を推奨した．しかしながら，彼自身も指摘しているように，レイヤー間のインターフェイス設計の難しさ，複雑なシステムへの拡張性が大きな問題として立ちはだかっている．

そもそもロボットの処理は，与えられたセンサ信号に対し，なんらかの目的達成のための適切なモータ信号を出力

すること，つまり，なんらかの基準の下でのセンサ–モータ間処理の「最適化」が目的であると考えられる．前述のように，「意識」を通しての理解が限られているとすれば，人間が理解できることを優先してしまうと，予想以上にロボットの機能を制限し，柔軟性を奪う結果となる．たとえば，ロボットの行動計画の際に良く出てくる「目標軌道」も，実はわれわれの理解のためのものであり，直接モータ信号を学習させた方がより柔軟な制御につながると筆者は考えている．人間が内部を手動で開発するのでなければ，人間による理解よりもシステムの最適化に重点を置くべきである．また，与えられたセンサ信号の下でより良い行動を得るためには，このセンサ信号を適切に認識すること，そして，必要な情報を記憶することも求められる．したがって，単なる「最適化」とは言え，その結果としてさまざまな機能が必要に応じて内部に創発する可能性を持っている．また，システムの「最適化」が目的であれば，人間が下手に手を出して，システムの自由度，柔軟性を奪ってしまうのではなく，システム全体が統一的な基準の下で調和を保ちながら「最適化」されることが望ましい．ただし，柔軟性を阻害しない初期値としての機能付与や学習過程の拘束の重要性を否定するものではない．

また，われわれは，現在得られたセンサ信号と全く同じセンサ信号を将来得ることは二度とない．にもかかわらず，多くの場合，過去の類似した状況での経験を活かして，適切に行動することができる．これは，人間の非常に優れた能力であり，人間のような知能を持ったロボットを開発する上で，必須の機能である．そこで鍵となるのが，「抽象化」とその抽象化された空間上での「汎化」である．

「抽象化」の厳密な定義は難しいが，重要な情報を抽出し，不要な情報を捨てて圧縮することと捉えることができるであろう．Brooks は「“抽象化” は知能の本質であり，解決が困難なところである」と述べている[3]．たとえば，ロボットにテニスボールを打ち返す動作を学習させる際にカメラ画像上のボールの位置や大きさ等の情報を与えようとわれわれは考える．これらの情報が重要であるということを発見することこそが知能の本質なのに，それを一方的に与えてしまっている．また，試合に勝つためには，実は「まわりの明るさ」の情報を考慮することが重要だったとしても，設計者がその情報を前処理の段階で捨ててしまっていては学習のしようもない．前述のように「無意識」の中で重要な情報が処理されているとすれば，設計者の「意識」が重要と判断した情報を抽出するのでは，本当に重要な情報を捨ててしまう可能性が非常に大きい．

「抽象化」での問題は，「何が重要な情報か」の判断基準をどうするかということである．たとえば，「圧縮前の情報を再現できる度合い」が基準として考えられる．砂時計型のニューラルネット[24] や主成分分析の利用もこの基準に基づいたものと考えることができる．しかし，これでは合目的性がなく，必ずシステムの目的に合致するとは限らない．上記の議論からすれば，この「抽象化」の基準も，システムの「最適化」の基準と一致していることが望ましい．

以上より，**人間が「理解」することよりも「最適化」を優先し，超並列で柔軟なシステムの全体を，統一した基準によって「最適化」させること**，そして，**人間の下手な干渉をできるだけ排除し，システム自身の「最適化」に任せること**．これが，本記事における 3 つ目のポイントである．

3. 強化学習とニューラルネット

以上のような議論から，筆者は，**図 1** のように，センサからモータまでを 1 つのニューラルネットで構成し，強化学習に基づいて生成した教師信号でそれを学習することを提案してきた[1),5)]．ニューラルネットは並列かつ柔軟なシステムである．ニューラルネットは学習のために教師信号を必要とするが，人間が教師信号を与えていては，ニューラルネットはそれを越える能力を学習できない．強化学習はそれを自律的に生成することができ，ニューラルネットは，与えられた教師信号を元に，システム全体を合目的的かつ柔軟に「最適化」することができる．そして，これによってわれわれの予想以上の能力を獲得することもある．

一方，強化学習は，通常，状態と行動のマッピングを学習する「行動の学習」と捉えられているが，ニューラルネットを用いることで，より良い行動を生成するために必要な，「認識」を始めとするさまざまな機能を，学習を通して獲得することが期待できる．また，リカレントニューラルネットを導入すれば，記憶が必要となる機能の創発も期待される．ニューラルネットの内部表現は，システム全体の目的に沿って抽象化された情報であると考えることができる．また，並列柔軟な学習システムにより，今までとは違った実世界への柔軟な対応も期待されるし，あらかじめ機能を用意しておく必要がないので，フレーム問題[4] に対しても根本的な解決になる可能性がある．さらに，高次機能についても，ニューラルネット内部が非常に柔軟に学習されるため，あらかじめ入出力を定めなくても必要に応じて機能が創発してくる可能性が期待される．ただ，われわれ人間の

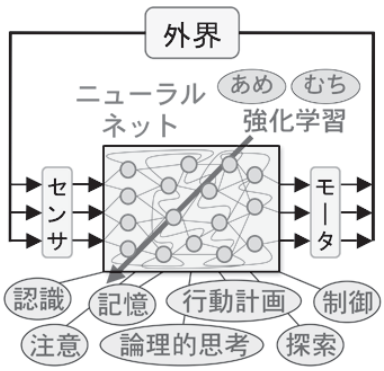

図 1　強化学習とニューラルネットを組み合わせた並列で柔軟，かつ合目的で汎用的な自律学習システム．必要に応じた機能創発が期待される．

持つ最も典型的な高次機能と言える「論理的思考」については，どういう必要性から創発するかの明確な筋道が現時点では見えていない．しかし，人間が実際に脳で行っていることなので，ニューラルネットによる学習も十分可能性はあるものと期待している．また，シンボルとパターンの処理の境界はなくなるため，この問題が解決すれば，シンボルグラウンディング問題[6]も本質的に解決されると期待される．さらに，「機能モジュール化」に変わる切り口として筆者が考える，「単純なものから複雑なものへと“成長発達するシステム”」という考え方に対しても相性が良い．

現在の強化学習におけるニューラルネットは，非線形関数近似器としての位置づけが強く，機能創発への期待はほとんどされていない．関数近似についても，当初は，倒立振子[7]や BackGammon[8]の学習と続いたが，学習の不安定性が指摘された[9]1995 年頃より勢いが鈍り，現在は代わりに正規化ガウシアン (NG) ネット[10]等がよく用いられる ([11] など)．強化学習のバイブル的存在の本[13]でも，ニューラルネットのことはあまり触れられておらず，その本の著者の 1 人である Sutton はインターネット上の強化学習 FAQ[14]において，様子見の必要性は指摘しているものの，ニューラルネットを用いてもあまり良い結果は得られておらず，使うのも難しいと否定的な見解を述べている．

しかし，ニューラルネットによる学習を見ていると，非常に自律的かつ柔軟に学習できることに驚く．各ニューロンは画一的な出力計算，学習 (重み値更新) 計算をしているにもかかわらず，中間層ニューロンが勝手に相互に役割分担をして，入出力関係の実現に必要な表現を合目的的に獲得していく様子は，単なる関数近似器としてではなく，機能創発，そして「知能」への可能性を感じさせる．

強化学習とニューラルネットを組み合わせると学習が不安定になるとの指摘[9]がある．NG ネットを含む RBF ネットやその拡張，および CMAC では[12]，テーブルルックアップのように連続値状態空間を局所的な状態に分割するため，局所的な状態ごとに学習でき，学習が安定になる．しかし，逆に，大域的な情報を内部に表現する手段がない上，内部表現は通常固定なので，ニューラルネットのように学習によって有用な大域的内部表現 (抽象表現) を獲得し，つぎのタスクの際にその空間上での汎化を利用した学習を行うことで，前のタスクで学習された知識を活用する[25]という重要な機能が実現できないという大きな問題点がある．

筆者らは，シグモイド関数ベースの通常のニューラルネットにおいても，入力信号を局所的な情報を表現する信号とすることで学習が安定することを示してきた[5), 15)]．また，視覚センサを始めとするセンサ信号は，もともと局所的な情報表現になっているため，センサ信号を単に直接ニューラルネットに入力すればよい．逆に，視覚センサ信号を，物体の位置等の連続値の情報に変換した後に入力すると，学習が不安定になることがあった．また，入力信号が局所的な情報表現であっても，その後の中間層でそれを統合した大域的な情報を表現することができるので，学習の安定性と，柔軟な内部表現の獲得の両者を同時に実現することができる．

一方，学習に重点を置いた最近の知能化の研究において，過去と現在のセンサ信号から自律的な学習が可能な「予測」の利用が注目されている[16)~20)]．しかしながら，非常に多数のセンサ信号の，将来の各ステップでの値をすべて予測することは不可能であろう．そう考えると，いつの何を予測させるかということが大きな問題となる．これは，前述の「抽象化」の議論と類似している．線形独立性から予測させる情報を発見する方法[21]も提案されている．しかし，「抽象化」の際と同様に，合目的性，システムの目的との整合性を考えると，やはり，「予測」も強化学習の中で扱うべきであると考えられる．そして，リカレントニューラルネットを用いて予測が必要となるタスクの学習を行なえば，内部に「予測」の機能も必要に応じて創発するのではないかと筆者は考えている．

つぎに，具体的に強化学習でニューラルネットを学習させる方法を述べる．actor-critic[22]の場合には，ニューラルネットの出力層に critic 用のニューロン 1 個 actor 用のニューロンをモータの数だけ用意する．そして，時刻 t でのセンサ信号ベクトル $\mathbf{S}_t$ を入力としたときの actor の出力ベクトル $\mathbf{O}a(\mathbf{S}_t)$ に試行錯誤の乱数ベクトル $\mathbf{rnd}_t$ を足したものを動作信号として実際に動作させ，その際新たに得られた報酬 r_{t+1}，および，センサ信号 $\mathbf{S}_{t+1}$ を入力して得られた critic の出力 $Oc(\mathbf{S}_{t+1})$ を用いて，critic および actor の出力ニューロンの教師信号 Tc と $\mathbf{T}a$ を，

$$Tc_t = Oc(\mathbf{S}_t) + \hat{r}_t = r_{t+1} + \gamma Oc(\mathbf{S}_{t+1}) \qquad (1)$$

$$\mathbf{T}a_t = \mathbf{O}a(\mathbf{S}_t) + \alpha \hat{r}_t \mathbf{rnd}_t \qquad (2)$$

$$\hat{r}_t = r_{t+1} + \gamma Oc(\mathbf{S}_{t+1}) - Oc(\mathbf{S}_t) \qquad (3)$$

と求める．ただし，$\hat{r}$ は TD 誤差，γ は割引率，α は定数である．そして，時刻 t でのセンサ信号 $\mathbf{S}_t$ を再入力してフォワード計算し，上記の教師信号でニューラルネットを学習させる[24]．一方，Q-learning[23]の場合は，行動と同数の出力ニューロンを設け，センサ信号ベクトル $\mathbf{S}_t$ を入力したときのそれぞれの出力 $O_a(\mathbf{S}_t)$ を各行動に対応した Q 値として用いる．そして，行動後の新しいセンサ信号 $\mathbf{S}_{t+1}$ を入力して得られた Q 値の最大値を用いて，時刻 t において選択された行動 a_t の出力ニューロンに対する教師信号を

$$T_{a_t,t} = r_{t+1} + \gamma(\max_a O_a(\mathbf{S}_{t+1})) \qquad (4)$$

と求める．ただし，O_a は行動 a に対応する出力である．そして，やはり時刻 t でのセンサ信号 $\mathbf{S}_t$ を再入力してフォワード計算をした後に，該当出力ニューロンのみ学習する．ただし，actor, critic, Q 値の値域と実際の出力ニューロンの値域がずれる場合は，適宜線形変換すればよい．このように，1 ステップさかのぼる学習の不自然さは残るものの，学習は非常に簡単で，汎用的なものである．

4. 学習の例

センサ–モータ間を単にニューラルネットでつなぎ，何の知識も与えずに強化学習で学習するだけでは，とてもではないが実世界での複雑なタスクの学習や高次機能の実現は無理との認識が一般的である．本章では，実世界への対応，記憶，抽象化，コミュニケーションに関する，簡単ではあるが今後の可能性を感じさせると筆者が感じる学習例を示すので，読者ご自身で今後の可能性を判断していただきたい．詳細は，各文献を参照されたい．また，以下の(*)の印の部分の関連した図，動画は，http://www.sice.or.jp/~journal/article/48-01-b.html に示す．

4.1 実世界に近い環境での柔軟な認識の学習[1), 26)]

2台の SONY 製犬型ロボット AIBO を用いて，2つの実験を行った．1つは，両 AIBO を向かい合わせに置き，片方の AIBO が首を動かし，正面に相手の AIBO を捉えた時に吠えると報酬が，そうでない時に吠えると罰がくるという学習をさせた[1)(*)]．もう1つは，片方の AIBO を歩かせ，相手の AIBO とキスをした時に報酬を与え，見失った時に罰を与えるという学習を行った[26)]．前者では，首は，5° ずつずらした離散的な 9 状態のうちの 1 つをとり，行動は，首を右回転，左回転，吠えるの 3 つとした．後者では，前進，右旋回，左旋回の 3 行動とし，相手の AIBO との距離，方向ともにさまざまな状態をとる．いずれも，カメラからの 52 × 40pixel の RGB カラー画像の 6240 個の信号を，各層のニューロン数が 6240-600-150-40-3 の 5 層のニューラルネットへ直接入力し，Q-learning で学習した(*)．また，いずれの場合も，照明条件や背景を変化させた．**図 2** に後者の場合のサンプル画像を示す．タスクの情報は一切与えられず，それほど簡単な問題ではないことがわかる．学習方法の詳細は文献を参照していただきたい．

学習に用いていないパターンでのテストを行ったところ，人間には遠く及ばないものの，学習当初はまったく成功していなかったものが，最終的には前者では 9 割程度，後者でも 8 割から 9 割程度の学習成功率となった．(*)(動画あり)．

最下層の中間層の 600 個のニューロンそれぞれへの入力信号からの結合の重み値の数は，入力画像の信号数と同じであるため，初期重み値からの変化量を線形変換して 1 つの画像として表現した．作成された 600 個の画像を見ると，首振りタスクの場合には，多くの場合，**図 3**(a) のように，画像中にポジまたはネガの AIBO の姿がはっきりと見られ，その位置はニューロンによって異なっていた(*)．また，1つの画像にポジとネガの両方の AIBO の姿が見えたものが多かった．これは，ニューロン間での自律的な役割分担を通し，限定された AIBO の位置に対する並列パターンマッチング，および，背景や明るさによらない認識を実現していると推測される．また，真ん中の中間層の表現を観察するため，最下層の中間層ニューロンの重み値の変化を，真ん中の中間層ニューロンへの重み値で重みづけし，その和

(a) (b) (c)

図 2 入力画像のサンプル．背景や照明条件によって画像が大きく変化していることがわかる．

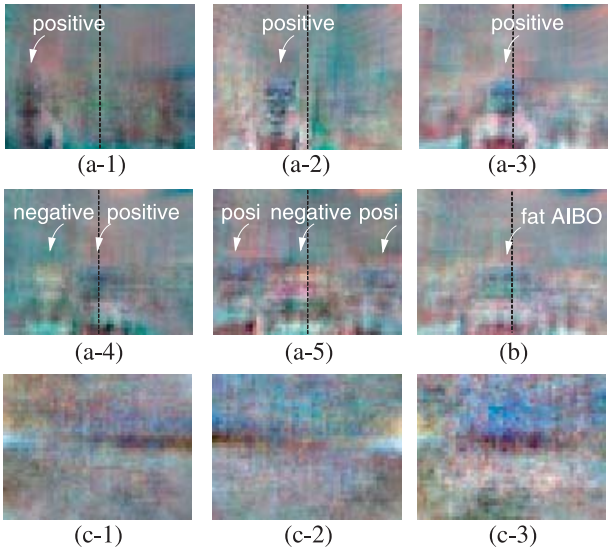

図 3 (a) 最下層の中間層のニューロンへの入力からの結合の重み値の学習による変化を画像化したもの．小さい白い矢印のところに **AIBO** やそのネガのイメージが見える．(b) 真ん中の中間層のニューロンの重み値の変化を画像化したもの．(a) (b) は首振りタスクの場合．(c) 歩行タスクの際の真ん中の中間層のニューロンの重み値変化のイメージ

を画像化したところ，図 3 (b) のように太って見える AIBO が多く観察された．これは，実ロボットでの実験のため，首の位置制御が必ずしも正確にできない影響を吸収する働きがあるのではないかと推測される．また，学習前後のニューラルネットワークを比較したところ，学習後には，明らかに，背景や明るさによらず，AIBO の存在に注目した内部表現が中間層の中に獲得されていることも確認できた[1)(*)]．

一方，歩行タスクの場合の，真ん中の中間層ニューロンでは，図 3 (c) のように，画像の中央の少し上部に横に伸びる白または黒の細長い帯があり，それが上下に重なっていたり，横方向に伸びている途中で白黒反転しているものが多く見られ，白黒の色反転の位置やその帯の上部の広がりが異なっていた．歩行タスクの場合，図 2 からもわかるように，相手の AIBO の見え方のバラエティが非常に多いので，このような形で背景や照明条件に影響されないで，AIBO の，特に顔の辺りが横方向および前後方向のどこにあるのかを検出していると推測される．

このように，学習によって，タスクに応じた柔軟な内部表現と適切な認識ができることを確認した．ただし，これ

らの解釈は，ニューラルネットの並列処理に対する一側面であり，実際の処理をすべて説明することは困難である．

4.2 記憶の学習と文脈依存行動の発現[27)]

つぎに記憶を必要とするタスクの学習をシミュレーションで行った結果を報告する．車輪型のロボットが，スイッチを踏んだ時に知覚される Flag1 の信号が 1 の場合にはゴール 1 へ，Flag2 の信号が 1 の場合にはゴール 2 へ到達すると報酬が得られる設定とした．ロボット，スイッチ，ゴールは毎回ランダムに置かれ，ロボットは，スイッチとゴールの角度と距離の情報，壁までの距離の連続値の情報，2 つのフラグの情報を得て，actor-critic タイプのリカレントニューラルネットで学習を行った．正しいゴールに到達した時の報酬，壁に衝突した時，スイッチを踏まずにゴールに来た時，間違ったゴールに来た時の罰以外は何も与えない．すると，ロボットは最初に必ずスイッチに向かい，そこで得られるフラグの信号にしたがって正しいゴールに向かうことができるようになった．学習後の中間層に，スイッチを踏んだ時に得られる Flag1 の情報を保持するもの，Flag2 の情報を保持するもの，どちらかのフラグが入ったことを保持するものの 3 種類のニューロンが観察された．さらに，Flag1 の情報を保持している複数のニューロンのうちの 1 つの値を強制的に下げても，すぐに元の値に戻ったことから連想記憶の機能も獲得したと考えられる．

つぎに，スイッチ上で Flag1 を知覚した後にゴール 1 に向かう途中の中間層ニューロンの値を保管した．そして，つぎの試行時に，Flag2 の方を 1 にセットし，**図 4** (a-1) のように，スタートから行動を開始した．そして，スイッチを踏んだ後にゴール 2 に向かう途中で，中間層ニューロンの値を先ほど保管した値と置き換えた．するとロボットは，向きを変えてゴール 1 に向かい，さらに，ゴール 1 で報酬が得られず試行が終了しないと，図 4 (a-2) のように，再びスイッチに戻り，フラグを確認した後，ゴール 2 に行って報酬を得た．この時の critic の出力と，Flag1 の情報を保持する中間層ニューロンの値の時間変化を図 4 (b) (c) に示す．ゴール 1 で試行が終了しないことで，critic の値，それから，値の入れ替えから大きな値を保持していた中間層ニューロンの値がともに大きく落ちており，予想外のことが起こると不安になって確認に戻る人間の姿と重なって興味深い．

4.3 合目的的な抽象的状態表現と知識転移の学習[28)](*)

エージェントが 2 種類のセンサ S1, S2 と 2 種類のモータ M1, M2 を持ち，使用するセンサとモータを 1 つずつ毎試行ランダムに選ぶ．そして，エージェントが動いて目標位置に到達した際に報酬を与えて学習させた．ニューラルネットには，S1 からと S2 からの両方の信号を入力し，選択されていないセンサからの信号は常に 0 とした．出力層には，M1, M2 用の 2 セットの actor-critic の出力を用意し，選択された方のモータ信号に基づいてエージェントを動かした．そして，S1-M1, S1-M2, S2-M1 の 3 つの組み合わせのみで学習をさせたところ，学習していない S2-M2 の組み合わせでもある程度正しい行動ができるようになった．ニューラルネットは入力が違っていても，教師信号が似ていれば，中間層表現が似てくるという性質がある[25)]．これにより，S1-M1 と S2-M1 の組み合わせでの学習を通して，S1 と S2 の別々のセンサを使った場合でも，状態が同じである場合には同じような内部表現となる．さらに，S1-M2 の学習によって，センサによらない状態に対する M2 の出力が学習されることで，S2-M2 の場合も学習することなく適切な行動が獲得された．センサ信号だけを見て抽象化の方法を決定する場合には，使用するセンサが違えば，それが同じ状態であることを認識することは不可能である．しかし，ここでは，強化学習による合目的的な抽象化により，使用するセンサによらない状態表現が可能になった．

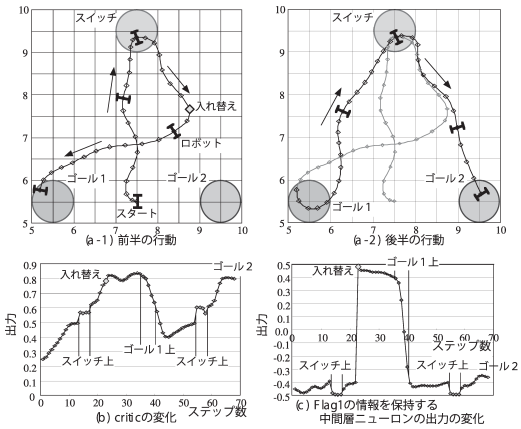

図 4　中間層ニューロンの出力を試行途中で入れ替えた場合の (a) ロボットの行動，(b) critic と (c) Flag1 の情報を保持する中間層ニューロンの出力の変化

4.4 コミュニケーションの学習と信号の二値化[29)](*)

典型的な高次機能の 1 つであるコミュニケーションの学習例について報告する．2 つのエージェントのうち，片方は知覚した情報を元に信号を発し，もう片方は，受け取った信号を元に行動を行うことができる．そして，後者が行動して目的を達成すると両者が報酬をもらえるという簡単なタスクの学習を行った．学習前には，送信側は何を送信するか，受信側も何が送られてくるのかまったく知らないにもかかわらず，学習後，両者の間でタスクに必要な情報を，その信号を使ってコミュニケーションするようになった．また，連続値の信号にノイズをのせると，学習による信号の二値化が観察され，離散情報表現であるシンボルの学習による発現の可能性を示唆した．また，リカレントニューラルネットの導入によって，引き続く 2 つの信号で必要な情報を表現し，コミュニケーションできるようになることや信号の二値化が促進されることも確認した．

5. 結論

高次機能実現のためには，従来の「機能モジュール化」の

考えから脱出し，脳のような並列で柔軟，かつ，合目的的で汎用的な自律学習システムを導入すること，そして，そのシステム全体を統一的な基準で「最適化」し，人間が下手に手を出さないことが重要であるとの考え方を解説した．そして，そのためには，センサからアクチュエータまでをニューラルネットでつなぎ，強化学習で学習させる方法がよいことを紹介し，簡単なタスクではあるが今までにない創発能力を示した学習例を紹介した．今後は，「論理的思考」が学習できるかどうかが大きな鍵となると筆者は考えている．最後に，このような方向性を持つ研究は，今後人間のコントロールが利かないロボット，システムが出現する可能性を秘めており，監視と議論の必要性を提起したい．

（2008 年 9 月 22 日受付）

参 考 文 献

1） 柴田克成，河野友彦：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，第 18 回インテリジェント・システム・シンポジウム講演論文集，145/150 (2008)
2） P. Johansson, L. Hall, S. Sikstrom and A. Olsson: Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task. Science, **310**, 116/119 (2005)
3） R. A. Brooks: Intelligence Without Representation. Artificial Intelligence, **47**, 139/159 (1991)
4） D. Dennett: Cognitive Wheels: The Frame Problem of AI, The Philosophy of Artificial Intelligence, Margaret A. Boden, Oxford University Press, 147/170 (1984)
5） 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，**37**–2, 168/177 (2001)
6） S. Harnad: Symbol Grounding Problem, Phisica D, **42**, 335/346 (1990)
7） C.W. Anderson: Strategy learning with multilayer connectionist representations, Proc. of the 4th Int'l Workshop on Machine Learning, 103/114 (1987)
8） G.J. Tesauro: TD-Gammon, a self-teaching backgammon program, achieves master-level play, Neural Computation, **6**–2, 215/219 (1992)
9） J.A. Boyan and A.W. Moore: Generalization in Reinforcement Learning, Advances in Neural Information Processing Systems 7, The MIT Press, 370/376 (1995)
10） J. Moody and C.J. Darken: Fast Learning in Networks of Locally-Tuned Processing Units, Neural Computation, **1**, 281/294 (1989)
11） 森本淳，銅谷賢治：強化学習を用いた高次元連続状態における系列運動学習：起き上がり運動の獲得，信学論，**J82-D-II**–11, 2118/2131 (1999)
12） R.S. Sutton: Generalization in reinforcement learning: Successful examples using sparse coarse coding, Advances in Neural Information Processing Systems 8, MIT Press, 1038/1044 (1996)
13） R.S. Sutton and A.G. Barto: Reinforcement Learning: An Introduction, A Bradford Book, The MIT Press (1998)
14） R.S. Sutton, ed.: Reinforcement Learning FAQ, http://www.cs.ualberta.ca/˜sutton/RL-FAQ.html (2001, 2004)
15） 柴田克成，前原伸一，伊藤宏司：Gauss-Sigmoid ニューラルネット，計測自動制御学会 システム・情報部門学術講演会 2002 講演論文集，467/472 (2002)
16） M.L. Littman, et al.: Predictive Representations of State, In Advances in Neural Information Processing Systems 14, MIT Press, 1555/1561 (2002)
17） J. Schmidhuber: Exploring the Predictable, Advances in Evolutionary Computing, Springer, 579/612 (2002)
18） J. Tani: Learning to generate articulated behavior through the bottom-up and the top-down interaction processes, Neural Networks, **16**–1, 11/23 (2003)
19） R.S. Sutton, E.J. Rafols and A. Koop: Temporal abstraction in temporal-difference networks, Advances in Neural Information Processing Systems 18 (2006)
20） P.-Y. Oudeyer, F. Kaplan and V.V. Hafner: Intrinsic Motivation Systems for Autonomous Mental Development, IEEE Trans. on Evolutionary Computation, **11**–1, 265/286 (2007)
21） P. McCracken and M. Bowling: Online discovery and learning of predictive state representations, In Advances in Neural Information Processing Systems 18 , 875/882 (2006)
22） A.G. Barto, et al.: Neuronlike adaptive elements can solve difficult learning control problems, IEEE Trans. on Systems, Man, and Cybernetics, **13**–5, 834/846 (1983)
23） C.J.C.H. Watkins: Learning from Delayed Rewards, PhD thesis, Cambridge University, Cambridge, England (1989)
24） D.E. Rumelhart, et al.: Learning Internal Representation by Error Propagation, Parallel Distributed Processing, MIT Press, **1**, 318/364 (1986)
25） 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，**43**–1, 54/63 (2007)
26） K. Shibata and T. Kawano: Learning of Action Generation from Raw Camera Images in a Real-World-like Environment by Simple Coupling of Reinforcement Learning and a Neural Network, to appear in Proc. of ICONIP 2008
27） H. Utsunomiya and K. Shibata: Contextual Behavior and Internal Representations Acquired by Reinforcement Learning with a Recurrent Neural Network in a Continuous State and Action Space Task, to appear in Proc. of ICONIP 2008
28） K. Shibata: Spatial Abstraction and Knowledge Transfer in Reinforcement Learning Using a Multi-Layer Neural Network, Proc. of Fifth Int'l Conf. on Development and Learning (2006)
29） K. Shibata: Discretization of Series of Communication Signals in Noisy Environment by Reinforcement Learning, Adaptive and Natural Computing Algorithms, Proc. of ICANNGA'05, 486/489 (2005)

［著 者 紹 介］

柴田克成（しばたかつなり） 君（正会員）

1989 年東京大学大学院工学系研究科機械工学専攻修士課程修了．89 年 (株) 日立製作所入社．92 年 10 月同社退職．93 年東京大学大学院工学系研究科先端学際工学専攻博士課程中退．同年東京大学先端科学技術研究センター助手．97 年日本学術振興会未来開拓学術研究推進プロジェクト研究員 (東京工業大学在籍)．2000 年大分大学工学部電気電子工学科講師．02 年より同助教授．04〜05 年 Alberta 大学客員教授．主として，ニューラルネットを用いた強化学習・自律学習システムの研究に従事．