G-014

DVB 強化学習を行うリカレントニューラルネットワークの構造獲得

Evolving Recurrent Neural Networks Trained by Direct-Vision-Based Reinforcement Learning

齋藤 成人 †　　服部 元信 ‡
Naruto Saitou　　Motonobu Hattori

1. はじめに

ロボットが人間のように多様に変化する実環境でも適切に行動できる知識を獲得するためには，さまざまな環境で自ら学習して，必要となる知識と機能を自律的に身につけるシステムが必要であると考えられる. このような複雑な情報処理を行う機構の一つとして，脳を模倣した情報処理方式であるニューラルネットワークの研究がなされてきた. その一つに，Direct-Vision-Based(DVB) 強化学習 [1] がある. この手法は，ニューラルネットワークの学習に強化学習を取り入れ，学習に必要な教師信号を自動的に生成し自律的な学習を可能にした. しかし，ネットワークの構造設計，複雑なタスクでの知識獲得の困難さなどに問題点もあった. また，ニューラルネットワークが様々な環境に適応できる知識を獲得する上で，ネットワークの構造や扱える情報は非常に重要であり，複雑な環境に適応するには，ネットワーク構造が時間情報を扱えることが望ましい. 時間情報を必要とする問題をニューラルネットワークに学習させる研究はすでに多く行われているが，それらの研究では設計者がそれぞれのタスクに適したニューラルネットワークの構造を予め試行錯誤的に設計している点で自律的であるとはいえない. また，リカレントニューラルネットワーク (RNN) の構造を遺伝的アルゴリズム (GA) で獲得している手法 [2] もあるが，DVB 強化学習を実装した研究は行われていない.

そこで，本研究では，GA に基づき，DVB 強化学習を用いて，RNN に時系列学習を要するタスクを自律的に学習させることを目的とする. RNN を用いることによって，時系列情報を含む複雑な環境を記憶し，GA によって，タスクに適した RNN 構造の自律的な獲得，DVB 強化学習によって自律的なタスクの学習を行うことが期待できる.

2.RNN の構造とコード化

RNN とは，過去の情報をフィードバックする結合をもつネットワークのことである. 本研究で採用する RNN 構造を図 1 に示す.

3 階層の階層型ニューラルネットワークに，過去の情報をフィードバックする過去入力層，過去中間層，過去出力層を新たに追加した. これらの層によって与えられる過去の情報によって，時系列による入出力の変化を記憶することが期待できる. また，様々な過去の情報をフィードバックする意図は，タスクによってどの過去の情報が重要であるかは違うことが推測されるからである.

また，GA に用いる個体 RNN を図 1 のようなコードによって表現した. 入力層ニューロン数 i，出力層ニューロン数 o を基本とし，中間層ニューロン数 h，過去入力層数 p，過去中間層数 b，過去出力層数 d は個体によって違い，可変とする. それぞれのニューロン同士の接続を，接続している場合は 1，接続していない場合を 0 として表現する. r, u, j, v, w はそれぞれの重みを表し，右下添字は接続の位置を表すニューロンの番号を示し，右上添字はその重みが何時刻前の情報 (値) との接続の重みであるかを示している. また，ネットワークとして成り立たせることと，DVB 強化学習を導入するため，入力層-中間層，中間層-出力層の接続は全て 1 とした.

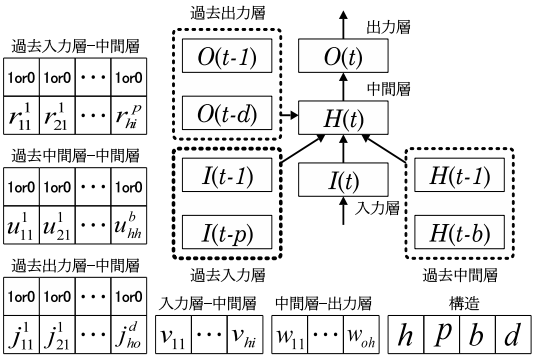

図 1: RNN の構造とコード化

3.GA による構造獲得

GA とは，生物の進化を模倣したアルゴリズムであり，交叉・突然変異・選択・淘汰を繰り返すことによって，優秀な個体を獲得する手法である. 本研究での交叉，突然変異，選択・淘汰の方法について以下で説明する.

3.1. 交叉

親個体群 P の中から，M 個のペアをランダムに選択し，それぞれのペアに対して交叉を行い，$M \times 2$ 個の子個体を生成する. 中間層ニューロン数，過去入力層数，過去中間層数，過去出力層数によって決定する構造の遺伝子は，交叉を行わず，それぞれの親から別々の子に引き継がれ，子のコードの長さが決定する. その後，両親の両方に存在する遺伝子は一様交叉による交叉を行い，片方の親にしか存在しない遺伝子は，構造を引き継いだ子にそのまま引き継がれる. 一様交叉を採用した理由は，構造の多様性を維持するためである.

3.2. 突然変異

交叉によって生成された子個体群 R の個体に対して，中間層ニューロン数，過去入力層数，過去中間層数，過去出力層数，接続をそれぞれ変異率 P_h，P_p，P_b，P_d，P_{code} で突然変異させた. 以下は，それぞれの突然変異が起こった場合の処理である.

中間層ニューロン数，過去入力層数，過去中間層数，過去出力層数の場合は，変異する個体の各数値を基準に，それぞれ $[-M_h, M_h]$，$[-M_p, M_p]$，$[-M_b, M_b]$，$[-M_d, M_d]$ の範囲で数値をランダムに決定し，変異させた. コードは子に存在する遺伝子は引き継ぎ，増加によって増えた元々存在しない遺伝子の接続，重みはランダムに設定した.

†山梨大学大学院医学工学総合教育部, 甲府市
‡山梨大学大学院医学工学総合研究部, 甲府市

接続の変異の場合は，要素ごとにそれぞれ P_{code} の確率で対立遺伝子に変異させた．

3.3. 選択・淘汰

選択・淘汰は親個体群 P と子個体群 R を合わせた全個体群に対して行った．全個体群に対して選択・淘汰を行う利点は，交叉や突然変異によって生まれた子個体の適応率が低い場合の解の悪化を防ぐことができるからである．また，選択・淘汰の基準である適応度はタスクの成功率とし，適応度が一番高い個体をエリート個体として選択し，その後，エリート個体を除く全個体群に対してルーレット選択を行い，局所解に陥ることを防いだ．

4. 実験タスク

時間情報が必要なタスクとして，エージェントがスタート位置から正しいルートを辿りゴールまで到達することを成功とするタスクを行った．まず，図 2 に示すマップを用意し，エージェントをスタート位置に配置する．エージェントは，[上，下，右，左] の 4 行動をとることができ，学習過程において正しいルートを辿った場合は各 step ごとに報酬を与え，それ以外のルートを辿った場合は，罰を与えエピソードを終了した．また，ニューラルネットワークの入力としてエージェントに与えられる情報は，周囲 8 マスの状態であり，障害物がある場合は 1 を，それ以外は 0 をそれぞれ入力値として与えた．そのため，例えば，図 2 の状態 1，2，3 のときに，それぞれエージェントに与えられる情報は全く同じであるため，正しい過去の情報を保持していない限り，状態 1，2，3 で間違った行動をとり，ルートから外れてしまいタスクを達成できないように設定した．

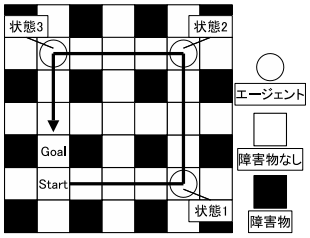

図 2: 時間情報を必要とするタスク

5. 実験結果

比較対象として，過去の入力情報を与える Time Dependent Neural Netwark(TDNN)，過去の中間層の情報を与える Elman 型 Neural Netwark(ENN)，Multi-Context Recurrent Neural Netwark(MCRNN)，過去の出力情報を与える Jordan 型 Neural Netwark(JNN) の 4 種類とした．それぞれのネットワークに対して中間層ニューロン数を {5,10,20,30,40}，過去層数を {1,2,3,4,5,10} とし合計 90 パターンの構造を用意した．TDNN と JNN はそれぞれ 30 パターン．ENN は過去中間層を 1 層利用するものなので，5 パターン．MCRNN は過去中間層を複数用意するネットワークであり，過去中間層数は 2〜10 層であるため，25 パターン．これが 90 パターンの内訳である．1 試行ごとのエピソード数を 10000 回とし，100 エピソードごとにテストを行い，タスク成功率が 100% になったら試行を終了し，各パターンをそれぞれ 100 試行行った．提案手法のパラメータを表 1 に，実験結果を表 2 に示す．表 2 の比較対象の結果は，各構造で一番良い結果を示したニューロン数，過去層数の組み合わせの結果である．また，表 3 は，提案手法によって獲得された個体において，タスクを成功した構造 (94 個体) の過去層の使用割合である．

表 1: 実験条件

パラメータ	値
世代数	100
エピソード回数/1 世代	100
親個体群 P	30
子個体群 R	40
P_h,M_h	0.25, 5
P_p, M_p	0.25, 2
P_b, M_b	0.25, 2
P_d, M_d	0.25, 2
P_{code}	0.01
報酬	0.9 ×step 数/Goal までの step 数
罰	−0.9

表 2: タスク成功回数とタスク成功率の平均

構造	タスク成功回数	タスク成功率 [%]
提案手法	94	99.1
TDNN	0	32.7
ENN	0	28.2
MCRNN	14	62.2
JNN	61	91.0

表 3: 成功した構造の過去層の使用割合

	過去入力層	過去中間層	過去出力層
割合 (個体数)[%]	13(12)	68(64)	85(80)

タスクの成功回数と成功率から任意に設定した一般的な RNN では獲得できなかった知識を同じ学習条件において，提案手法では獲得できている点から，タスクに適した構造を獲得できたといえる．また，JNN のタスクの成功回数が他比較対象と比べて良い結果を示していることからもわかる通り，今回のタスクは過去の行動 (出力) が重要な情報であり，提案手法では，過去出力層の使用割合が高いことからも，適切な構造を獲得できているといえる．

6. まとめ

本研究では，GA に基づき，DVB 強化学習を用いて，RNN に時系列学習を要するタスクを自律的に学習させることを目的とし，時間情報を必要とするタスクに対する知識の獲得，獲得された構造の検証を行った．実験結果より，リカレントニューラルネットワークを用いることによって，時系列による環境の変化を記憶し，遺伝的アルゴリズムによって，タスクに適したリカレントニューラルネットワーク構造を自律的に獲得，Direct-Vision-Based 強化学習によって自律的なタスクの学習を行うことを示した．

参考文献

[1] 柴田 克成, 岡部 洋一, 伊藤 広司,“ニューラルネットワークを用いた Direct-Vision-Based 強化学習–センサからモータまで–,” 計測自動制御学会論文集, Vol.37, No.2, pp168–177, 2001.

[2] M.Delgado, M.P.Cuellar, M.C.Pegalajar,“Multi-objective Hybrid Optimization and Training of Recurrent Neural Networks,”*IEEE Transactions on Systems, Man, and Cybernetics–PartB:Cybernetics*, vol.38, No.2, pp.381–403, 2008.