RNNを用いた強化学習による センサ信号の時間変化を表すコミュニケーションの創発

大分大学　○朱祺　柴田克成

Emergence of Communication that Represents Time Change of Sensor Signals by Reinforcement Learning Using a RNN

∗Qi Zhu and Katsunari Shibata, Oita University

Abstract− In order that robots acquire “real coummunication” by solving the “Symbol Grounding Problem”, not only exchange of words, but also other functions such as recognition should be learned togeteher, the authors think. Previous researches have shown that coummunication was acquired autonomously and comprehensively through reinforcement learning with a neural network. In this paper, the possibility of communication representing time change of sensor signals by reinforcement learning using a recurrent neural network(RNN) was examined. In the employed task, a sender agent needs to control its visual field, recognize the object movement from the time change of a vision-like simple sensor signals, and then send signals to a receiver. The receiver needs to identify the coummunication signals and make the judgment at an approriate timing. As the results, firstly, it was shown that communication representing the time change of the sensor signals emerges, while a preliminary learning for sender is required before the learning. Secondly, to make communication learning successfully, it was suggested that learning parameters such as learning rate and initial weight range should be set individually between sender and receiver.

Key Words: Reinforcement Learing, Recurrent Neural Network(RNN), Symbol Grounding problem, Robot Coummunication

1　まえがき

現在のコミュニケーションロボットは、一見人間と適切に会話しているように見える。しかし、実際に会話を聞いていると、ロボットは言葉の意味を理解して喋っているとは考えられず、依然として「シンボルグラウンディング問題」は解決されていないように見える。その原因は、ロボットがやり取りしている言葉だけに注目し、設計者によって予め与えられた手順に従って音を発しているだけだからと考えられる。ロボットが真のコミュニケーションを獲得するには、言葉のやり取りの部分だけを取り出して学ぶだけではできず、センサ信号や相手からの信号をもとに、現在の状況を適切に認識するところから、信号を発生させるまでのすべての機能を総合的に学習していく必要があると考えられる。また、そもそもコミュニケーションを行うためには、無数のセンサ情報が得られる中で「どのような情報を相手に伝えるか」を自ら判断することが非常に重要であることは言うまでもない。

当研究室では、センサからモータまでを並列処理の学習が可能なニューラルネットで構成して強化学習で学習することで、膨大なセンサ情報を入力とし、機能ごとに分けることなく、処理全体を自律的かつ総合的に学習することが可能となり、その結果、ニューラルネット内部に報酬を得て罰を避けるために必要な機能が事前に与えることなく創発することを示してきた [2]。

先行研究 [3] では、このような枠組みをコミュニケーション学習にも適用した。視覚を持つ送信者がコミュニケーション信号を発生し、視覚を持たない受信者がその信号を受けてロボットの動作を生成し、ロボットがゴールに着くと両者が報酬を得られるタスクを学習させた。そして、単に報酬や罰だけに基づく学習によって、送信者が多くの入力信号からロボットの位置情報を抽出し、受信者に送ることができ、一方、受信者は受け取った情報から適切にロボットを動かして、ゴールに到達するようになり、送信者から受信者へ目的に沿った適切なコミュニケーションを獲得できることを確認した。

しかし、人間のコミュニケーションに動詞があるように、ロボット間でコミュニケーションさせる場合も物の名前や状態だけでなく、状態の変化を伝える必要もあると考えられる。例えば、情報としてボールの画像があって、それについてコミュニケーションを行うとする。「これはボールだ」「どこにボールがあるか」といった情報の伝達だけではなく、「ボールがどう動いているか」といった動きの情報を伝達することが必要な場合もあるだろう。コミュニケーションの学習は、送信者、受信者がともに適切な信号生成や処理をした時だけしか報酬が得られないため、ただでさえ学習が難しいが、現時点だけの情報を処理してできるコミュニケーションとは違い、動きの情報を伝えるためには過去の時系列情報を処理する必要があり、学習がさらに困難になることが予想される。

本研究では、報酬を得るためにはセンサ信号の時間変化の情報の伝達が必要なタスクを、記憶能力のあるリカレントネット (RNN) を用いた強化学習で学習させ、送信者がセンサ信号の時間変化を観察し、コミュニケーションによって何を伝達すべきかを判断し、受信者が受け取った情報をどういう行動に反映させればよいかを判断するなどの機能を自律的かつ総合的に獲得させ、センサ信号の時間変化を表すコミュニケーションが創発するかどうかを確認する。

2　タスクおよび学習方法

本研究で行った、センサ信号の時間変化の情報を信号として伝達する必要があるタスクの様子を Fig.1 に

示す。離散空間である 5×5 の格子上において、物体が毎試行開始時に中心部の 3×3 のどこかにランダムに出現し、その試行中、上、下、左、右の中からランダムに選んだ運動を行う。各試行開始時には送信者のセンサは中央の 3×3 の部分にあるとし、各マスに物体があれば 1、なければ 0 の計 9 個の信号を RNN に入力する。送信者は、4 つの信号のうちどれかを 1 にするか、4 つの信号をすべて 0 にしてセンサを上、下、左、右、4 方向のうちどれかに動かすか、または動かさないかの計 9 個の行動からボルツマン選択によって 1 つを選択する。受信者は、信号を受け取り、上、下、左、右、4 つの判断のうちどれかを下すか、待機して何もしないかの計 5 個の行動から ϵ-greedy 選択によって行動選択を行う。ϵ は式 (1) のように試行回数 trial によって徐々に減小させた。ただし、ランダムで選択するとき「待機」判断が選ばれる確率を 0.6, 他の方向の判断が選ばれる確率をそれぞれ 0.1 とした。

$$\epsilon = 0.3 * 5^{-trial/2000000} \tag{1}$$

送信者と受信者は別々の RNN を持ち、送信者は、センサ信号入力の 9 個と 1 ステップ前の自身のセンサの動き (上、下、左、右、待機) の 5 個、計 14 個を入力とする。出力は 9 個用意し、センサを待機、上、下、左、右に動かすための 5 個と信号出力用に 4 個の Q 値として用いた。受信者は、送信者からの 4 個の信号を入力し、待機、上、下、左、右の判断を下すために 5 個の出力を Q 値として用いた。ここで、ニューロンの出力関数は-0.5 から 0.5 の範囲のシグモイド関数を用い、その値に 0.5 を加えて Q 値とした。物体の動きと送信者のセンサの動きはともに 1 ステップあたり 1 マスずつとし、受信者が待機以外の判断を出したときに、それがの正解か不正解かによって両者に報酬またには罰を与えて、次の試行に移る。また 1 試行で 5 ステップ経過しても、報酬 (罰) を得られない場合はその試行を打ち切る。

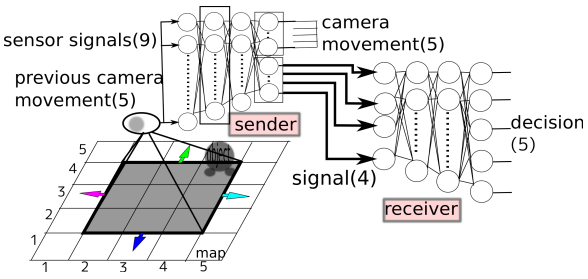

Fig. 1: Communication Task

学習の流れを以下に示す。

1. 送信者が物体の状態 S_t を観測し、自身のニューラルネットに入力する。
2. 送信者のニューラルネットの出力を計算し、行動 A_t を選択し、信号の出力、センサの動作を行う
3. 発生した信号を受信者のニューラルネットに入力する。
4. 受信者のニューラルネットの出力を計算し、行動 A_t を決定する。
5. (受信者が判断を下した場合) 正解なら教師信号を $r_t = 0.9$(報酬に相当)、不正解なら $r_t = 0.1$(罰に相当) を両者に与え、BPTT 法で両者の RNN を学習させ、当試行を終了する。
6. (受信者が「待機」を選択した場合) 送信者は $t+1$ の時の物体の状態 S_{t+1} を入力, 最大 Q 値の行動 (センサ移動か信号送るか) をとり、受信者が信号を受け、出力を生成する。式 (2) により教師信号 $T_{A_t,t}$ を作り (割引率 γ は 0.7 とする)、BPTT 法で両者の RNN を学習させる。時間ステップを t から $t+1$ に進めて 1 に戻る。

$$T_{A_t,t} = \gamma max_A Q(S_{t+1}, A) \tag{2}$$

その他のパラメータ設定を Table 1 に示す。

Table 1: Parameter setting

Number of neurons in each layer		
Sender: 14(input)-30-20-10-9(output)		
Receiver: 4(input)-30-20-10-5(output)		
Traced back time for BPTT	Sender	5
	Receiver	1
Inital connection weights in sender		
Input-Hidden1 (in Preliminary learning)		-0.1∼0.1
Hidden1-Hidden2, Hidden2-Hidden3 (in Preliminary learning)		-0.2∼0.2,-0.5∼0.5
Hidden3-Output		-4∼4
Inital connection weights in receiver		
Input-Hidden1		-0.1∼0.1
Hidden1-Hidden2, Hidden2-Hidden3		-0.2∼0.2,-0.5∼0.5
Hidden3-Output		-2∼2
Self-feedback, Other-feedback		4, 0.5
Learning rate		
For feedback connections		0.0125
For other connections	Sender	0.01
	Receiver	0.3

3 学習結果

3.1 送信者の予習におけるセンサ移動の獲得と物体動作の学習

送受信者の双方を 0 からコミュニケーションの学習をさせたところ、学習できなかった。先行研究 [3] においても、コミュニケーションを 0 から学習させることは難しく、送信者がまず自分で直接ロボットを動かす学習をすることで、学習が大幅に加速することが示されている。認識とコミュニケーションを同時に学習させる場合、学習初期に偶然必要な情報が信号として出力されて報酬を得られる可能性が極めて小さくなって、学習困難であることを示唆しており、本タスクではさらにセンサを適切に動かすことも求められる。そこで、ここでも、まず送信者のみで、Fig.2 に示すように、送信者センサ移動部はそのままとし、信号の出力の代わりに、本来受信者の出力である判断のための出力を設けて、物体の動きの認識を予習として学習させた。

物体の動作を認識するためには, まず、送信者が物体の位置によって自分の持つセンサを適切に動かす学習

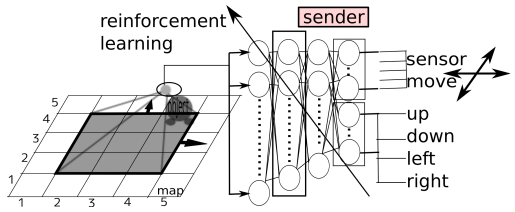

Fig. 2: Preliminary learning in the sender

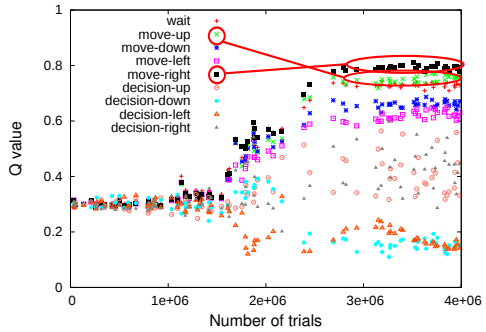

Fig. 3: Q-value changes in the sender (1st step, preliminary learning)

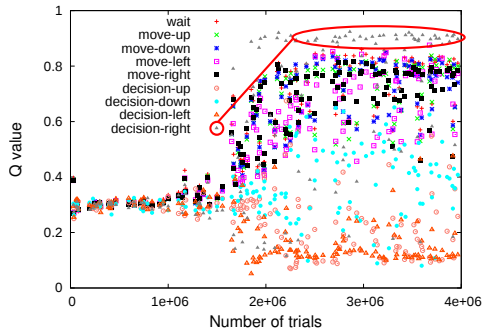

Fig. 4: Q-value changes in the sender (2nd step, preliminary learning)

が必要である。例えば最初のステップに物体がセンサの端の格子 (4,4) に出現し、右か上に移動した場合、送信者がセンサを右か上に動かさないと、いずれの場合も物体を見失ってしまうため、正しく動作を認識できない。Fig.3 は物体が格子 (4,4) に出現し、まだ右移動をしていない時の送信者各行動の Q 値が学習の進行 (試行回数) に対してどのように変化したかを表したグラフである。学習初期、送信者のすべての行動に対する Q 値が 0.3 付近に集まったが、学習によって、センサを右に動かす行動 (■) と上に動かす行動 (×) に対する Q 値が他の行動の Q 値より大きくなっており、送信者がセンサを正確に動かすことができるようになったと言える。

次に、物体が格子 (4,4) に出現して右移動をして、センサが右に動いた後の、学習の進行に対する送信者各行動の Q 値の変化を Fig.4 に表す。学習によって、右の判断の Q 値 (▲) が他の行動の Q 値より大きくなり、物体の右運動に対して, 正しく判断下すことができるようになった。同様に, 物体の上、下、左方向への運動を行った場合にも対して、正しい判断を下すことができるようになった。

Fig.5 は、予習後の物体が上、下、左、右に運動をする場合のそれぞれについて、送信者のある中間層ニューロンの出力の変化を示す。Fig.5 において、9 本の線は物体の初期位置が違う 9 個の場合の送信者の出力の変化を表している。この中間層ニューロン出力は物体の初期位置によらず、物体の左運動の場合だけ 2 ステップ目に出力が大幅に増加することがわかる。同様に、物体の上、下、右動作のそれぞれに反応する中間層ニューロンが存在することも確認した。

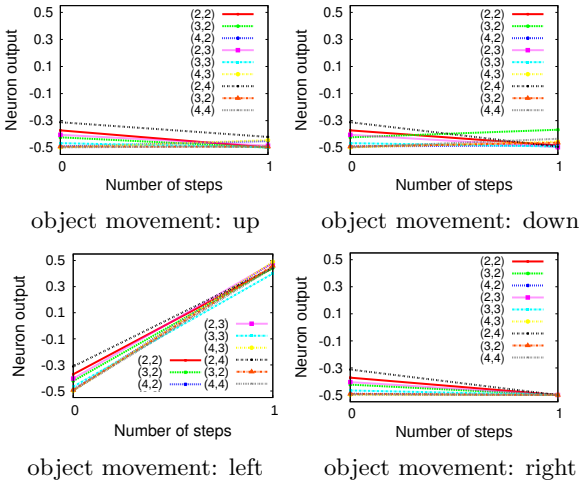

Fig. 5: A hidden neuron which learned to responds to the left movement of the object

3.2 コミュニケーションの学習

コミュニケーションの学習を行う際、送信者は物体の動きの認識を行う代わりに受信者に送る信号を生成する。したがって、基本的には予習した送信者の重み値をそのまま使い、出力の信号発生部のみ中間層ニューロンとの間をランダムな重み値で結合した。受信者と送信者間のコミュニケーションを学習させた結果、300 万試行で、センサ信号の時間変化を表すコミュニケーションを学習できた。

物体が右運動を行った際の最初のステップでの受信者の各判断に対する Q 値の学習による変化を Fig.6 に示す。学習初期は、受信者は信号がどういう意味を表してるのかを認識できず、すべての Q 値が 0.4 弱の付近に集まった。学習によって、送信者がセンサを動かすとき, 受信者の待機の Q 値 (+) が他の行動の Q 値より高くなり、次の信号が来るまで待機ことができるようになった。

Fig.7 は、次のステップに受信者の各判断に対する Q 値の学習による変化を示す。グラフに示しているように学習によって、受信者の右判断を下す Q 値 (■) が他の行動の Q 値より高くなり、送信者とコミュニケーションをとることによって正しい判断を下すことができるようになった。

4 学習パラメータの影響

コミュニケーションの学習は、送信者と受信者がともに適切な選択をしなければ報酬がもらえないので、学習が難しい。そこで、学習がうまく進むためには学習

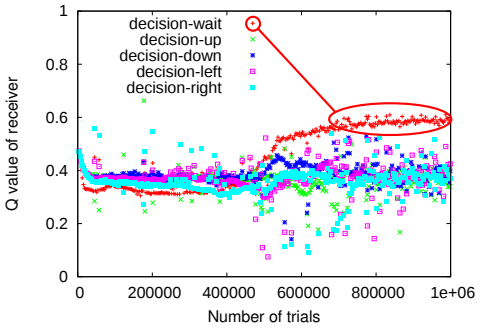

Fig. 6: Q-value changes in the receiver (1st step, communication learning)

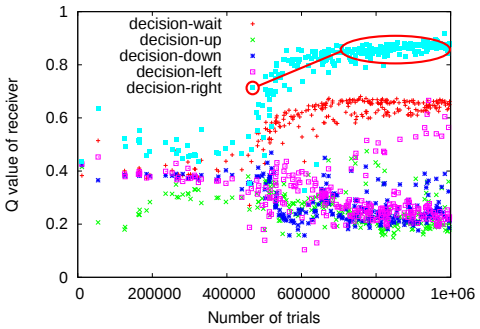

Fig. 7: Q-value changes in the receiver (2nd step, communication learning)

時のパラメータをどのように設定にすると良いかを調べた。送信者の行動選択温度、送信者と受信者の学習率、出力層の初期重み値設定時の乱数の値域を変化させた時に、乱数系列を変えた 100 試行の平均学習成功率がどのように変化したかをそれぞれ Table 2、Table 3、Table 4 に示す。

まず送信者のボルツマン選択での行動選択時の温度は低い方が成功率が高いが、温度が低過ぎて greedy 選択に近くなると、逆に成功率が落ちていることがわかる。つまり、送信者がわずかにランダム要素をもつ場合のコミュニケーション学習成功率が高いことが分かった。

次に、送信者と受信者の学習係数の影響を見ると、受信者と比較して送信者の学習係数がかなり小さいときに、学習成功率が高いことがわかる。

予習後、送信者の中間層にはコミュニケーションに必要な情報が表現されている。一方、受信者は、送信者の出力である信号を入力として学習する。したがって、同一入力時の送信者の出力が高い温度や大きな学習係数によって変化しない方が受信者にとって学習しやすかったと考えられる。そこで、先行研究 [4] においても、コミュニケーション時、送信者が温度と学習係数を小さくするほうが学習成功率が高いという結果が報告されている。

最後に、初期重み値用乱数の値域の影響について、受信者より送信者の値域を大きくする方が学習成功率が高い傾向が見られ、送信者が-5~5, 受信者が-2~2 のときに学習成功率が最も高かった。これは、送信者の出力層重み値にバラつきを持たせて、出力ニューロン (信号) ごとに表現する情報をある程度変えることで、各信号間での表現の分化を加速したためと考えられる。

Table 2: Success rate for temperature of action selection during learning

Temperature	greedy	0.01	0.05	0.1
Success rate	0.26	0.51	<u>0.80</u>	0.07

Table 3: Success rate for the combination of sender's and receiver's learning rate

learning rate		sender				
		0.01	0.05	0.1	0.2	0.5
receiver	0.1	0.61	0.66	0.54	0.40	0.07
	0.2	0.74	0.77	0.70	0.48	0.10
	0.3	<u>0.80</u>	0.66	0.63	0.35	0.07
	0.5	0.61	0.60	0.46	0.24	0.00

Table 4: Success rate for the combination of sender's and receiver's initial weight

Initial weight		sender				
		-2~2	-3~3	-4~4	-5~5	-6~6
receiver	-1~1	0.43	0.72	0.75	0.78	0.72
	-2~2	0.47	0.71	0.80	<u>0.88</u>	0.85
	-3~3	0.39	0.64	0.68	0.68	0.70
	-4~4	0.27	0.53	0.60	0.57	0.65
	-5~5	0.17	0.42	0.56	0.51	0.55

5 まとめ

本論文では、リカレントネット (RNN) を用いた強化学習を行うことで、センサ信号の時間変化を表すコミュニケーションを学習させた。送信側が物体の動きの認識を予習することでコミュニケーションの学習が可能であることを確認した。そして、学習パラメータの影響について、送信側の学習係数を小さく、初期重み値を大きくし、温度を低くすることで、学習成功率が高くなることを確認した。

しかしながら、現時点では 2 ステップに渡るだけの離散センサ信号の時間変化を表す簡単な情報のコミュニケーションであり、より長いステップに渡る変化、動詞と呼べるような情報のコミュニケーションを学習させることは今後の課題である。

参考文献

[1] Harnad, S. The Symbol Grounding Problem. Physica D 42, pp335-346, 1990

[2] 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, 2009.1

[3] 笹原冬月, 柴田克成. ニューラルネットと強化学習を用いた音声によるコミュニケーションの自律学習, 第 28 回 SICE 九州支部学術講演会, pp85-88, 2009

[4] 仲西賢展, 柴田克成.Q 学習に基づく一方向コミュニケーション学習に置ける学習効率化手法の提案, システム・情報部門学術講演会講演論文集, pp157-162, 2004