3R-08

深層強化学習によるシグナル伝達を介したマルチエージェントフォーメーションの形成

野口渉† 飯塚博幸† 山本雅人†
北海道大学 大学院情報科学研究科†

１．はじめに

マルチエージェントシステム(MAS)は，限られた機能しかもたないエージェント群が協調的に振舞うことで複雑なタスクを達成可能な点で優れたシステムである．協調的動作が求められるタスクとしてエージェントが整列して図形を形作るフォーメーションタスクがある．Rubenstein らは，1000 体からなる実ロボットを用いて複雑な図形を形作るシステムを構築した[1]．このシステムにおいては，基点となる少数の個体からの距離を個体間で伝播することでフォーメーションを実現しているが，基点となるエージェントが取り除かれた途端に機能しなくなってしまう点で望ましくない.自然界の鳥や魚などは，各個体がコミュニケーションを介して群における自身の位置や役割を決定していると考えられるが，複雑なフォーメーションを形成するためには，複雑なシグナルを用いた個体間の情報伝達が必要となる.

個体間のコミュニケーションを用いた MAS は多数提案されているが、Foerster らは，Q 学習の枠組みにおいて，エージェント間でのシグナル伝達に誤差の逆伝播を組み込むことで，効率的にコミュニケーションを発達させる Differentiable Inter-Agent Learning (DIAL) を提案した [2]. DIAL は近年注目される深層強化学習に基づいている．Foerster らは，タスク達成に必要とされる各個体の情報が変化しないタスクにおいて DIAL が有効に働くことを示したが，フォーメーションのような個体の位置が変化するタスクに対する DIAL の有効性は検証されていない．また，Foerster らの実験においては，全個体が大域的に通信可能な MAS を用いたが，各個体の位置が未知であるフォーメーションタスクにおいては，大域的な通信が有効であるとは限らない.

本研究では，DIAL がフォーメーションタスクにも有効であることを示し，局所的にシグナル伝達が可能な場合と大域的にシグナル伝達が可能な場合を比較する.

Multi-agent formation using inter-agent signaling with deep reinforcement learning
Wataru Noguchi†, Hiroyuki Iizuka† and Masahito Yamamoto†
†Graduate School of Information Science and Technology Hokkaido University

２．深層強化学習によるシグナル伝達

2.1 Deep Q-Networks (DQN)

DQN は，Q 学習における Q 値を出力する関数として，ディープニューラルネットワーク(DNN)を用いたシステムである[3]．DQN は，DNN により複雑な環境情報を認識することで適切な行動選択の学習を可能にしている．過去の経験を反復して学習する experience replay や，目標の Q 値を出力する target network を用いることも学習を安定化させる要因である．また，再帰的ニューラルネットワーク(RNN)を用いて，部分観測可能な問題に対して適用可能にしたモデルも提案されている.

2.2 Differentiable Inter-Agent Learning (DIAL)

DIAL は，RNN を用いた DQN の一種であり，時刻tにおける各エージェントaの Q 値は，センサ情報s_t^a，他エージェントからのシグナル$m_{t-1}^{a'}$，ネットワークの内部状態h_{t-1}^a，前ステップの行動u_{t-1}^aによって決定される．Foerster らは各エージェントに割り当てた識別番号も入力として用いているが，本研究ではホモジニアスなエージェントを仮定するため識別番号は用いない．DIAL において，エージェントのコントローラネットワーク(C-Net)は，以下の(1)-(3)式により Q 値(Q)と他エージェントへのシグナルm_t^aを出力する.

$$h_t^a = \text{C-Net}^h\left(s_t^a, m_{t-1}^{a'}, h_{t-1}^a, u_{t-1}^a\right), \quad (1)$$
$$Q = f^Q(h_t^a), m_t^a = f^m(h_t^a). \quad (2)$$
$$m_t^a = f^m(h_t^a). \quad (3)$$

ここで，C-Net^hは内部状態h_t^aをもつ RNN ユニット，f^Qは線形変換，f^mは線形変換とtanh関数による変換である．学習は Backpropagation Through Time (BPTT)を用いた Q 値の最適化によって行う．m_t^aは次ステップにおいてC-Netに入力され，逆伝播される Q 値に対する誤差をもとにm_t^aを最適化する．本研究では，入力を変換する ELU ユニットと，ELU ユニットの出力を受け取り Q 値とシグナルを出力する GRU ユニットにより，C-Net^hを構成する．ELU のサイズは 16，GRU のサイズは 16 とする．また，シグナルm_t^aの次元は 4 とする.

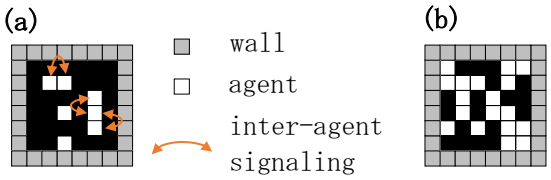

図 1: (a)環境とエージェント.
(b)目標フォーメーション

3. フォーメーションタスク

DIAL を適用するマルチエージェントフォーメーションにおける環境を設計する. 環境は図1(a)に示す2次元グリッド空間である. エージェントの行動は上下左右の4方向への移動と停止の5 種類であり, 環境中の壁と隣接するエージェントによってエージェントの行動は制限される. エージェントは上下左右の隣接するマスの状態を知覚可能なセンサを搭載する. センサの値は, エージェントが 1, 壁が-1, 空のマスは 0 とする. 次章では, 図 1(b)に示すフォーメーションを目標としてエージェントを学習する. このフォーメーションは, 6×6の各マスに対して確率0.5でランダムに目標位置（埋めるべきマス）として決定した複雑なフォーメーションである. 個体数は目標フォーメーションに必要な数と同数とする.

4. 実験

前章で設計したフォーメーションタスクに対するエージェントの行動を DIAL により学習するまた, 次の3つの条件それぞれに対して学習を行ない, 結果を比較する. (1)各個体は上下左右に隣接する個体に対してシグナル伝達可能（局所的シグナル伝達）. (2)各個体が他の全個体に対してシグナル伝達可能（大域的シグナル伝達）. (3)シグナル伝達なし.

4.1 実験設定

報酬は目標のフォーメーションと一致したマスの数を最大の一致数で割ることで[0,1]の区間に正規化した値とする. 一致したマスにいるエージェントにのみ報酬を与え, それ以外のエージェントに対する報酬は0とする. 100 ステップを1エピソードとし, エージェントの位置は各エピソードの始めにランダムに初期化する. 探索にあたって ε-greedy 方策を適用する. ε の初期値は 1.0 とし, 1 エピソードごとに 0.0001 だけ, 0.1 まで減少させる. コントローラは 1 エピソードごと 10 回学習する. 学習バッチサイズは 10 とし, replay memory から 10 ステップの部分エピソードをランダムにバッチ数分だけ取り出して学習に用いる.

4.2 実験結果と考察

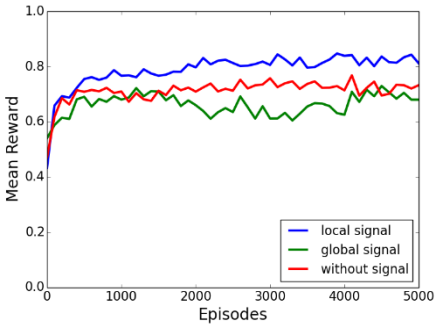

図 2: 平均獲得報酬の推移

図 2 に, ε=0 とした場合の 10 エピソードでの平均獲得報酬を学習エピソードごとに示す. 局所的シグナル伝達条件が最も獲得報酬が大きくなっており, 各個体の局所的センサ情報のみからでは適切な行動が不明な場合にも, 各個体のフォーメーションにおける位置や周辺の埋めるべきマスといった有用な情報がシグナルを介して伝達されることで, より正確なフォーメーション形成が可能となったと考えられる. 一方, 大域的シグナル条件は, シグナル伝達なし条件とほぼ同程度の獲得報酬となった.これは, 大域的にシグナル伝達することにより, 群全体で交換されうる情報量は局所的シグナル伝達の場合よりも増加する一方で, センサで観測できない個体から伝達される情報によって, 隣接する個体からのシグナルを有効に用いることが難しくなったためであると考えられる.

5. まとめ

本研究では, MAS のフォーメーションタスクに対して DIAL を適用することでシグナルを介した協調的フォーメーション形成が可能となることを示した. また, フォーメーションタスクに対しては, 大域的なシグナル伝達よりも局所的シグナル伝達の方が協調的動作の獲得が容易であることを示した. 今後の課題としては, より複雑なフォーメーションや, 連続空間でのフォーメーションへの DIAL 学習の適用があげられる.

参考文献

[1] Rubenstein, M., Cornejo, A., and Nagpal, R.: Programmable self-assembly in a thousand-robot swarm, Science, Vol. 345(6198), pp.795-799 (2014).

[2] Foerster, J. N., Assael, Y. M., de Freitas, N., et al.: Learning to Communicate with Deep Multi-Agent Reinforcement Learning. arXiv preprint arXiv:1605.06676 (2016).

[3] Mnih, V., Kavukcuoglu, K., Silver, D., et al.: Human-level control through deep reinforcement learning. Nature, Vol. 518(7540), pp. 529-533 (2015).