2G1-2

人とロボットの協働に向けた意図の生成と共有のモデル化

Model of generating and sharing intention: toward human-robot cooperation

グェン　アン　トゥアン　　日永田智絵　　長井隆行
Tuan Anh Nguyen　　Chie Hieida　　Takayuki Nagai

電気通信大学
The University of Electro-Communications

We normally tend to think that as human we act as if we have intention to achieve a goal. However, in the situation of cooperating with other people, it is difficult to coming up with a mechanism of predicting and adjusting each other's intention when we think that goal comes from top-down. Thinking that intention comes from context may help to solve that problem. In this study, we attempt to model an action learning mechanism that generates intention using Deep Q-Network and Recurrent Neural Network.

1. はじめに

近年，ロボットと人間がコミュニケーションする機会が増えている．この流れの中では，ロボットと人間の社会的インタラクションに関する様々な課題が表面化することが予想される．中でも大切なのは，ロボットが人間の行動に応じた適切な振舞いができるのかという課題である．我々は日常生活において他者を観察し，無意識にその裏にある意図を読みながら行動している．そうすることで，他者の意図を考慮して自分の行動を選択することができる．同様に，ロボットが人間と円滑にインタラクションできるためには，相手の意図を推定し自身の行動を調整する能力が必要である．

一方で神経科学の研究は，人間は目標達成という意図によって行動するという直感に反し，意図が行動の結果であることを示唆している [Libet 83, Haggard 05]．そもそも，目標や意図がトップダウンにあるとすれば，それはどこからきてどのように調整され得るのかが問題となる．また，意図から行動を決めると考えると，現状からその目標を達成するまでの行動を事前にプランニングすることが必要である．しかしインタラクションの場面では，他者の行動によってプランニングした結果と実際の状況とにずれが生じる可能性が高い．毎回誤差が生じると，それを修正するためにプランニングを繰り返す必要がある．また，誤差が出た状況から目標までのプランをいつも作成できるとは限らない．そのため，意図は事前に決められた目標により生成されるという考え方では計算コストも非常に高くなり，お互いの意図を予測し調整するようなメカニズムを考えることが困難である．そして，相手の意図を推定するメカニズムが自身の内部メカニズムに依拠しているのであれば，逆に自身の意図を自身で推定している可能性は高いのではないかとも考えられる．

そこで本研究では，トップダウンに目標や意図を持つのではなく，コンテキストによって意図が生み出されると考える．ここでコンテキストとは，これまでの行動の流れを意味する．エージェントはこの流れの中で自分の行動を受動的に決定し，この決定から自身の目標や意図を後付けで推定している．そして，あたかも目標や意図に沿って行動を決定したと感じる．本研究における仮説は，こうした受動性こそが他者とのインタラクションにおける意図の推定や目標の共有という感覚を生み，最終的に協調に至る基本的なメカニズムであるというものである．

連絡先: グェンアントゥアン、電気通信大学知能機械工学科、
n.a.tuan@apple.ee.uec.ac.jp

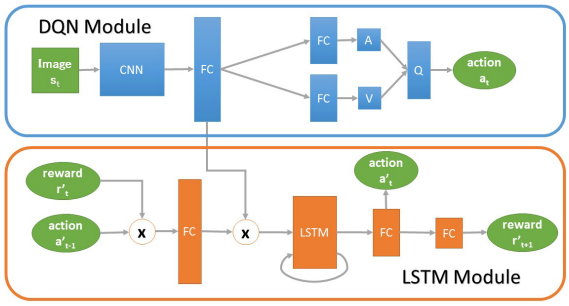


図 1: 提案モデルの概要図.

2. 提案モデル

上述の仮説をリカレントニューラルネットワーク (RNN) を使って言い換えれば，RNN に表現されたコンテキストを意図と感じ，その RNN におけるコンテキストから予測される未来が目標であるということになる．この仕組みを実現するために必要なのは，モデルフリーな強化学習と，モデルベースの RNN による行動学習の二つである．まずエージェントには具体的な目標が存在しないため，報酬を未来に渡って最大化する強化学習によって試行錯誤的に行動する．そして，行動することによって得る情報を RNN を用いてモデル化することで，行動のパターンとそれに応じた環境の変化，報酬の予測ができるようになる．この RNN のコンテキストが意図に対応し，RNN のコンテキストによって予測される未来が目標となる．ここでは，Dueling Deep Q-network (DQN) [Wang 16] と Long Short-Term Memory (LSTM) [Hochreiter 97] を組み合わせた意図と目標を生み出す行動学習モデルを提案する．

提案するモデルの全体像を図 1 に示す．ここで，DQN モジュールへの入力は環境の状態であり，一般には画像が用いられる．タスクに必要な視覚情報の特徴が，畳込みニューラルネットワーク (CNN) により抽出される．その後，いくつかの全結合層 (FC) の非線形変換によって，選択可能な行動の価値 (Q 値) [Watkins 92] が出力される．エージェントは与えられ

たタスクを遂行するために状態を観察し，Q 値に基づいて逐次行動を選択する．これを一つのステップと呼ぶ．また，初期状態から完了状態，又は失敗による終了状態まで一連の試行をエピソードと呼ぶ．DQN は，状態入力に応じて適切な行動を出力することを学習できるが，同じ状態でもコンテキストによって取るべき行動が違う場合に問題がある．例えば家のドアの前に立っている人は，今帰ってきたばかりなのか，家から出たばかりなのかというコンテキストにより次に取るべき行動が違う．しかし，強化学習ではこのような場面を区別できない．そこで，DQN によって行動した経験の全てのエピソードを保存する．これはいわゆるエピソード記憶である．そして，保存されたエピソードデータを用いて LSTM モジュールの学習を行う．LSTM モジュールに現状の画像の特徴を抽出した CNN の出力及び前ステップの行動や報酬を入力し，次のステップの行動と報酬を正しく予測するように教師あり学習を行う．LSTM は各エピソードのステップ間の連鎖を学習するので，同じ状態でもコンテキストに応じて取るべき行動が正しく選択できる．こうすることで，ある初期状態からタスク完成までを一連の行動として，予測に基づきながら実行することができる．

3. 実験

3.1 積み木タスク

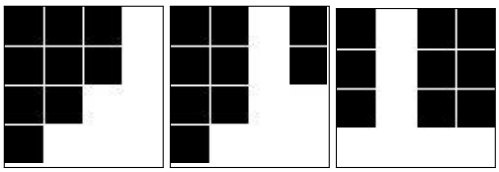
実環境での人間同士のインタラクションを簡素化したタスクとして，エージェント 2 体が協調できる積み木タスクを用いる．積み木を置く又は取り除く動作で，事前に決められたいくつかの目標形状を完成するタスクである．基本的なルールは以下となる．

- エージェントは，1 ステップに一度ずつ交互に積み木を置くまたは取り除く動作を行う．
- 事前に用意した目標形状をいずれか完成したらタスク終了とする．
- ある一定数以上のステップを繰り返しても目標形状を完成できない場合，タスク失敗と判断しタスクを終了する．

本実験で用いた目標形状を図 2 に示す．ただしこの目標形状をエージェントは知ることはできず，報酬によってのみ行動を決定する．このタスクで選択可能な行動は，ボード上の位置の 16 箇所に対応する行動とする．行動 1 とは，一番左上の位置を選ぶことに相当し，そこから左から右，上から下に番号をつける．積み木がない位置を選択した場合は積み木を置く動作とし，積み木がすでにある位置を選択すると取り除く動作とする．積み木は空中に浮いた状態で置くことはできず，下に積み木がない場所を選択した場合は，何も起こらずに次のエージェントの番となる．報酬 r は 5 段階で与えられる．最大 $r = 40$ はタスク完了時の報酬であり，その他は行動の良さによって $r = -4$，$r = -1$，$r = 1$，$r = 2$ として報酬を与える．

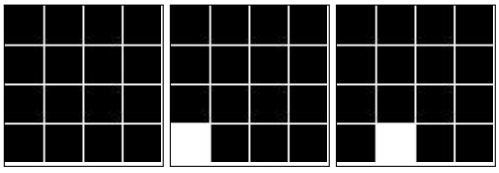
3.2 各エージェントの学習

エージェント 2 体は協調タスクの前に，それぞれが行動を学習する必要がある．各目標形状に対して，DQN は 15000 エピソード学習を行う．各エピソードでは 25 ステップを最大ステップとし，それを超える時をタスク失敗とする．入力画像は 84×84 のボード画像である．また各エピソードの初期状態を，図 3 に示す 5 つの状態からランダムに選択する．DQN は 3 つのゴールに対して，約 8000 エピソードから目標形状を完成することができるようになった．例として，図 4 に目標形状 3 における学習の様子を示す．図 4 の左は，学習エピソード数

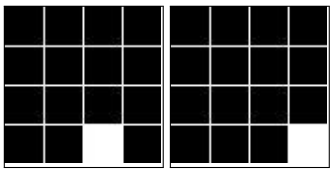


(a) No1 (b) No2 (c) No3

図 2: 目標形状.



(a) No1 (b) No2 (c) No3



(d) No4 (e) No5

図 3: 初期状態セット

に対するステップ数 (積み木を積む回数) を，右はエピソード毎の報酬をプロットしたものである．これより，最小の手順で目標の積み木を積むように学習されていることが分かる． 次に，DQN の学習中に保存されたエピソード記憶（合計 45000 エピソード）を用いて，5 エポックまで LSTM の学習を行った．学習後，環境の初期状態と目標形状をランダムに生成した 100 エピソードに対してエージェントがどのように振舞うかを確認したところ，すべての目標形状を完成することができた．次に，エージェントに予測した行動と報酬に従い行動させた．つまりエージェントは，環境からの報酬ではなく予測した報酬に基づいて行動する．すると，初期状態 1, 2 からスタートするエピソードでは目標形状 3 を作り上げる．初期状態 3, 5 のエピソードの場合，目標形状 1 を完成させた．また，初期状態 4 からは目標形状 2 を作るようになった．図 5 に示すのは各初期状態のエピソードにおいて，最初のステップから完成するまでの LSTM のコンテキスト出力ベクトルを PCA（主成分分析）により 2 次元に圧縮した結果である．目標形状 1 を作り上げる 2 つのエピソードでは LSTM の出力が似たような経路である一方，目標形状 2 や 3 の場合には少し違うパターンが見られる．ここでは，このコンテキストの軌道が意図を表しており，そのたどり着く先が目標を表現していると考える．図 5 にはさらに，破線で表現した目標形状 1 を作る場合のコンテキストの軌道と，対応した積み木の各ステップを示している．本稿では，以降このエージェントを，エージェント 1 と呼ぶ．

また図 6 は，異なる初期値から学習したもう一体のエージェント 2 について，同様にコンテキストの軌道を示したものである．エージェント 2 は，目標形状 1 と 2 を作るように学習されたことが分かる．

3.3 2 体による積み木タスク

3.2 で述べたエージェント 1 とエージェント 2 が，交互に積み木を積む実験を行った．既に述べたように，各エージェントは独立に学習しているため，ネットワークには異なるパターン

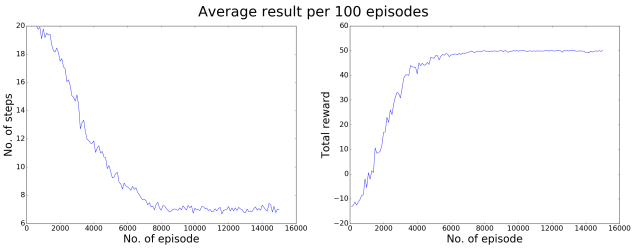


図 4: DQN の学習結果.

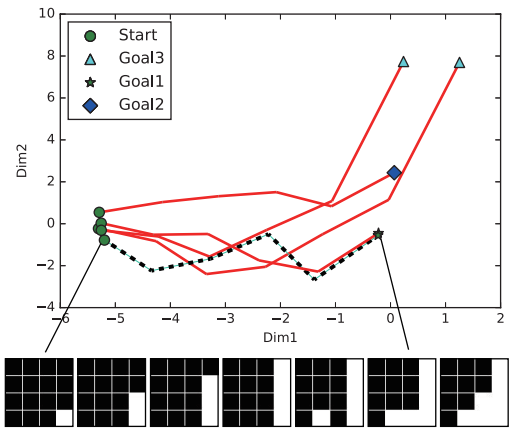


図 5: 初期状態の異なるエピソードにおける LSTM の出力の変化 (エージェント 1). 下段の積み木は，黒の破線で示したコンテキストの軌道の際に，エージェントが行った積み木の状態を順に並べたものである.

が学習されている．また，2 体はそれぞれ他のエージェントの存在については考えておらず，あくまでも積み木の状態のみを観測して行動する.

2 体は交互に，LSTM の予測行動結果に基づき行動する．各エピソードにおいて，いずれかのエージェントがタスク完了の報酬を予測した場合にそのエピソードを終了とする．ただし，コンテキストを様々なパターンに変化させるために各エージェントは，ϵ-greed 法 ($\epsilon = 0.1$) で定めた確率でランダムな行動を取るように設定した．これは，エージェント自身の探索や，環境のノイズを表現するものである.

図 7 に初期状態 5 から，エージェント 1 を先行として 2 体が行動した結果を示す．図 7(c)(d) において，白がエージェント 1 の積み木，灰色はエージェント 2 の積み木を表している．図 7(c) のパターンでは，目標形状 1 が作られていることが分かる．この例では，まずエージェント 1 が右下に積み木を積んだ後に，エージェント 2 が左から 2 番目に積み木を積んでいる．このエージェント 2 の行動が，エージェント 1 の意図を変化させ，本来であれば目標形状 1 を作る流れを変えている．(a) の左側の黒丸における分岐がこれを示しているが，エージェント 2 がいない状況では，右側の細い線に進み最終的に目標形状 1 を作る流れが学習されている．さらに (a) の右側の黒丸における分岐は，エージェント 2 の行動によって起きており，(c) の場合は目標形状 1 を作る流れに，(d) の場合は目標形状 2 を作る流れに引き込まれている．一方エージェント 2 では，(b) の破線黒丸で分岐が起きている．これは，エージェント 1 がエージェント 2 の行動によって目標形状 1 もしくは目標形状 2 に引き込まれ，その後にとった行動に影響を受けることで，目標形状 1 を目指す (c) の流れか目標形状 2 を目指す (d) の流れに分岐している.

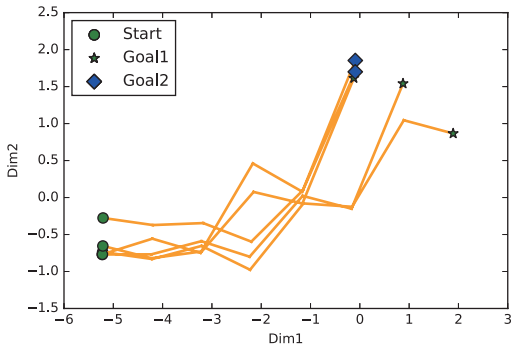


図 6: 初期状態の異なるエピソードにおける LSTM の出力の変化 (エージェント 2).

これらを見ると，2 体は相手の行動や意図を予測することなく，単にコンテキストに従って行動するだけで，目標形状が作られていることが分かる．そしてこの振る舞いだけを第三者的に見れば，2 者のエージェントが協調してある目標形状を作り上げているように見える.

当然，場合によっては振動してしまいうまく目標形状が作られない状況も存在する．各初期状態において 100 エピソードずつ交互に行動した結果，各目標形状がどれぐらいの割合で作られたかを，図 8 に示す． この図より，かなりの割合で目標形状が 2 体によって達成されていることが分かる．この割合は，エージェントの学習や ϵ の値によっても変化する．特に，ϵ によるランダム性が重要であると考えられる．ϵ の値はある意味，自身の意図にどれくらい固執するかを表現しており，値がゼロの場合は振動した場合に抜け出すことができない.

一方この実験では，最大のターン数を決めているため，そのターン数内に目標形状が作られなかった場合を失敗と定義している．失敗として作られた形状を，図 9 にいくつか示す．これらは，目標形状とは異なる形状が作られた結果であると捉えることもできる．つまり，目標形状という意味では失敗であるが，協調することで新たな形を創りだすという大きな意味があると考えられる．そのためには，こうしたインタラクション中の行動をさらに LSTM で学習する必要がある.

さらには，自他認知のメカニズムを取り入れることで，自分の意図や目標を推定するメカニズムを用いて，他者の意図や目標を予測することが可能であると考えられる．こうした点を，今後の課題とする.

4. 結論

提案したモデルは，コンテキストに応じて行動することによってある種の目標とそれを達成する意図を持つように振る舞うことを示す結果を得た．これは神経科学の解釈と同様に，行動する結果としての意図を生み出す可能性を示唆していると考えられる．今後の課題として，このような受動的な意図を持つエージェント 2 体がより複雑なタスクを交互に行うことで，どのような振る舞いを見せるのかを検討することが挙げられる．また，インタラクション中の行動も LSTM で学習し，新たな意図やゴールが生成されるかどうかを確かめたいと考えている．最終的には，自他認知のメカニズムを取り入れることで，インタラクションにおける目標や意図の共有を説明するモデルを構築したいと考えている.

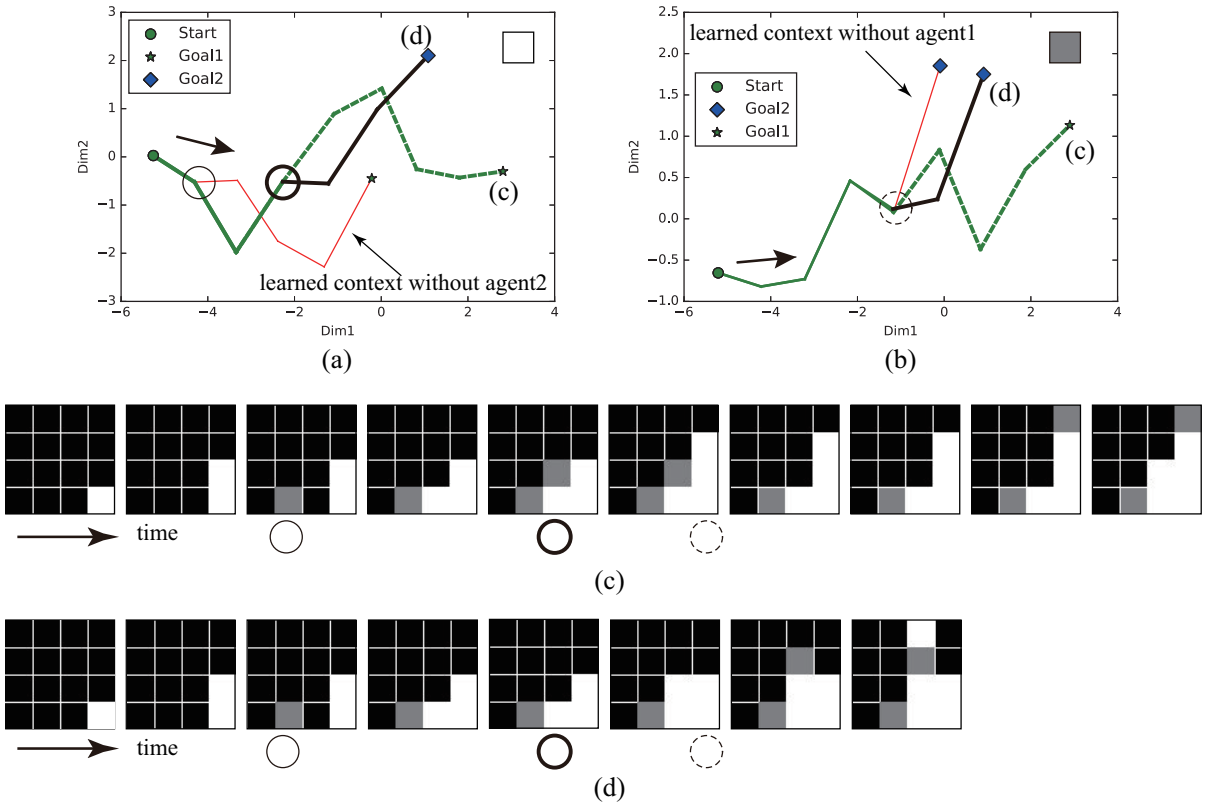


図 7: エージェント 2 体が交互に積み木タスクを行った結果．(a) エージェント 1 のコンテキスト PCA 空間，(b) エージェント 2 のコンテキスト PCA 空間，(c) 目標形状 1 が作られた事例，(d) 目標形状 2 が作られた事例．

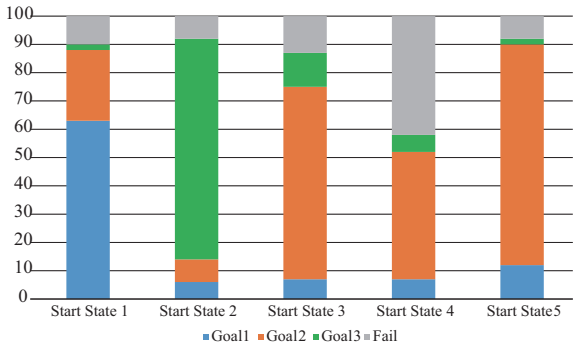


図 8: エージェント 2 体によって作られた目標形状の割合．

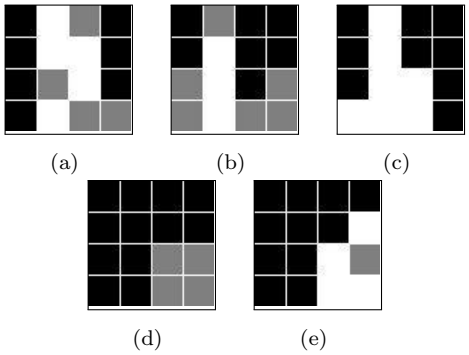


図 9: 失敗した時の形状

謝辞

本研究は JST CREST，新学術領域「認知的インタラクションデザイン学」の助成を受けたものです．ここに感謝の意を表する．

参考文献

[Libet 83] Libet B, Gleason CA, Wright EW, Pearl DK.: Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential), The unconscious initiation of a freely voluntary act. Brain 106, 623-642 (1983)

[Haggard 05] Haggard, P.: Conscious intention and motor cognition, TRENDS in Cognitive Sciences 9(6):290-295 (2005)

[Wang 16] Z.Wang, N.de Freitas, M.Lanctot: Dueling Network Architectures for Deep Reinforcement Learning, arXIv, 1511.06581 (2016)

[Hochreiter 97] Sepp Hochreiter and Jurgen Schmidhuber.: Long Short-Term Memory. Neural Comput. 9, 8, 1735-1780 (1997)

[Watkins 92] CJCH Watkins, P Dayan: Q-learning, Machine learning (1992)