3R-06

深層強化学習エージェントの自己モデルによる意図の解釈

福地庸介[1]　大澤正彦[1]　岨野太一[1]　山川宏[2]　今井 倫太[1]

[1] 慶應義塾大学
[2] 株式会社ドワンゴ　ドワンゴ人工知能研究所

1　はじめに

人間やエージェントが他者と協調して作業するためには，それぞれが互いの意図を推定し，推定結果に応じて行動する必要がある．他者の意図の推定には，環境や文脈といった相手を取り巻く情報や意図に直接関係する行動の情報だけでなく，表情，言語といった相手自身の発する付加的情報も重要と考えられる．

しかし深層強化学習によって行動獲得するエージェントは付加的な情報を生成する能力を持たない．また外部から行動の方策はわからない．そのため他者はエージェントの入出力情報のみから相手の意図を推定しなければならない．しかし，複雑な入出力を扱うエージェントに対して意図推定を行うことは困難である．

そこで本稿では，意図に直接関連する行動を出力する制御モデルと，制御モデルの意図を説明する自己モデルの 2 つを持つ Intention Expressing(IE) エージェントを提案する．

2　表現の意味の推定

相手が用いている表現手段を説明に利用することは，相手の説明理解に直結する．しかしエージェントが相手が使う表現を利用して自身の意図を説明するためには，表現の意味がエージェント内部に定義されている必要がある．これは例えば，すべてのエージェントに対して表現の定義をあらかじめ与えれば実現できる．

一方本稿では，エージェントが事前に表現の定義を共有するのではなく，他者とのインタラクションから獲得する方法を考える．意味定義を事前に固定しないので，環境と行動の関係を説明する表現を動的に獲得できると期待できる．具体的には，行動獲得の際に表現を戦略として与え，エージェントはこの際に表現の意味を推定する．表現の意味とエージェントの行動の対応づけは，与えられた表現が累積報酬の最大化を目的にした戦略であるという仮定の下で，学習で得られる報酬の情報から可能だと考える．つまり，ある戦略が与えられた中で，報酬が最大となる行動が，その表現の意味だと推定する．そしてエージェントは，獲得した表現をベースとして自身の行動の意図を説明する．

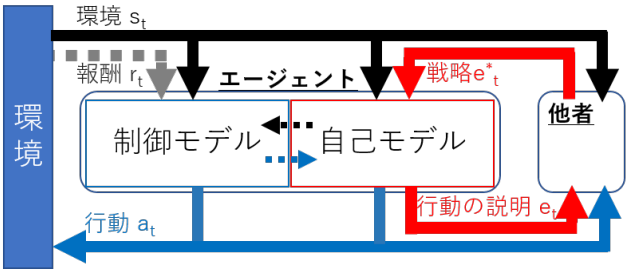

図 1: システム構成図　実線は他者から可視な情報，破線は不可視な情報

3　制御モデルと自己モデルの獲得

3.1　IE エージェントのシステム

図 1 にシステムと情報の流れを示す．ここでの他者とは，すでに自己モデルを獲得した IE エージェントを指す。IE エージェントは制御モデルと自己モデルで構成される．制御モデルは時刻 t における環境 s_t を入力とし，環境からの報酬 r_t をもとに累積報酬の最大化を目的として自身の行動 a_t を決定する Deep Q Network[1] を持つ．自己モデルは，制御モデルと s_t をもとに制御モデルの行動意図を解釈する．また，他者から与えられる戦略 e^* をもとに説明表現の獲得する．さらに，獲得した表現を用いて他者に向けた行動の意図の説明 e_t を出力する．本稿において e，e^* は，32bit の二進数である．エージェントはまず制御モデルによる行動獲得と自己モデルによる e^*_t の意味推定を同時に行う．ついで，e^* をベースに自己モデルが自身の行動に対する表現 e を獲得する．

3.2　行動獲得と戦略表現の推定

まず行動獲得と他者から与えられる戦略の意味推定を同時に行う．他者は毎時刻エージェントに戦略 e^*_t を与える．制御モデルは s_t を入力とする深層強化学習器 DQN_C を持つが，自己モデルも e^*_t の意味を推定する

Intention interpretation of a Deep Reinforcement Learning agent with self-model
[1] Keio University
[2] DWANGO Co., ltd Dwango Artificial Intelligence Laboratotry

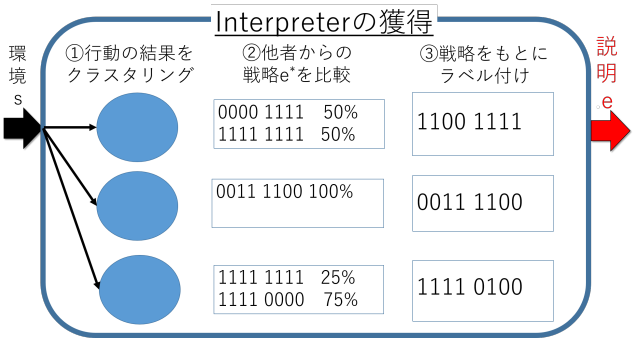

図 2: Interpreter の獲得

ため s_t, e_t^* の両方を入力とする学習器 DQN_S を持つ．DQN_S は，e_t^* が累積報酬の最大化を目的としているという仮定のもとに学習することで，e_t^* の意味内容を推定する．また行動獲得の際 ε -greedy 法を用いるが，greedy 方策の際 DQN_C が出力する行動と DQN_S が出力する行動を混合することで，制御モデルに e_t^* に沿った行動を誘発させる．最終的に制御モデルの行動 a_t は，e_t^* によく従っていると考える．

3.3 出力する説明表現の獲得

制御モデルの獲得の後，自己モデルは自身の行動を説明するための表現を獲得する．自己モデルは，行動による結果からそれを引き起こした行動の意図を解釈して説明する *Interpreter* と，ある時刻における行動による，次の時刻での環境の変化を予測する *Predictor* を用いて制御モデルの意図の説明をする．

s_t における制御モデルの行動の意図は，生成された行動の結果変化する環境 s_{t+n} から解釈できると考える．また行動獲得後，行動と戦略の対応関係から制御モデルが出力する行動 a_t は e_t^* によってよく説明されていると考える．そこで e_t^* をベースに s_{t+n} から a_t の意図の説明 e_t を出力する *Interpreter* を用意する．*Interpreter* 獲得の流れを図 2 に示す．まず s_{t+n} に対して多変数解析を行い，クラスタリングする．生成されたクラスタの単位で s_{t+n} を引き起こした a_t の説明 e_t を与える．各クラスタの環境に対しそれを引き起こした行動を行った時刻 t に与えられた戦略 e_t^* が存在する．クラスタを引き起こした行動群は，与えられた戦略群の意味内容を混合した意図を持っていたと考える．そこで各クラスタを実現する行動の説明として，行動の際に与えられてきた戦略の混合を利用する．具体的には与えられた戦略群の各 bit が 1 の確率を計算し，それに従って出力する説明 e の各 bit を決定する．以上の操作で与えられた戦略 e^* から自身の行動 a_t を説明するラベル e_t を得る．

また環境 s_t における自身の行動 a_t から環境 $\bar{s}_{t+1}$ を予測する *Predictor* を，教師あり学習により獲得する．*Predictor* が予測した $\bar{s}_{t+1}$ について制御モデルは自身の行動 $\bar{a}_{t+1}$ を決定でき，そこから $\bar{s}_{t+2}$ も予測できる．このようにを繰り返すことで，環境の変化 $\bar{s}_{t+n}$ が予測される．*Predictor* が $\bar{s}_{t+n}$ から行動の意図を解釈することで，エージェントは行動の段階で自身の行動の意図を説明できる．

4 実験

提案の検証のため，以下の二点を評価する実験を行う．IE エージェントの出力する説明が (1) 自身の行動と対応していること (2) 他者 (他のエージェント) から理解可能であること．今回の方法で出力される説明が以上の二点を満たしていれば，より多くの報酬を得ているエージェントによる自身の行動の説明は有力な戦略とみなせ，他エージェントの行動獲得を加速させることができると考える．そこで獲得された自己モデルが出力する説明を新たに行動獲得する他のエージェントに与え，学習の加速から出力された説明を評価する．

エージェントが操作する環境には，Lunar-Lander v2 というゲームを利用する．ロケットを地面に対し軟着陸して静止させることを目的とするゲームで，エージェントは 4 つのコマンド (ロケットの左/右/下からのジェット噴射，何もしない) でロケットの姿勢を制御する．初期状態として，人手により (左/右に行く, 姿勢を立て直す, 落下スピードを上げる, 落下スピードを下げる, 静止する)6 種類の戦略の組み合わせを出力する自己モデルを持つ IE エージェントを用意した．まず提案した手法で IE エージェントを複数用意し，そのうちより多くの報酬を得ているエージェント数体を評価対象とする．そして選んだエージェントが出力する意図の説明の有無による学習完了までの時間を比較する．

5 おわりに

本稿では，自身の行動決定モデルの意図を解釈し，外部に説明する IE エージェントの構築法を提案した．今後は 3.3 章で述べた実験を行いこれを評価する．

参考文献

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostro- vski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning, *Nature*, Vol. 518, No. 7540, pp. 529533 (2015)