人工知能学会研究会資料
SIG-AGI-005-08

深層強化学習エージェントの自己モデル獲得と行動目標説明表現の生成

Self-model acquisition and target explanation of a deep reinforcement learning agent

福地庸介[1]* 大澤正彦[1] 岨野太一[1] 山川宏[23] 今井 倫太[1]
Yosuke Fukuchi[1] Masahiko Osawa[1] Taichi Sono[1] Hiroshi Yamakawa[23] Michita Imai[1]

[1] 慶應義塾大学
[1] Keio University
[2] 株式会社ドワンゴ　ドワンゴ人工知能研究所
[2] DWANGO Co., ltd Dwango Artificial Intelligence Laboratotry
[3] NPO 法人 全脳アーキテクチャ・イニシアティブ
[3] The Whole Brain Architecture Initiative, a specified non-profit organization

Abstract: In teamwork, communication has an important role in understanding co-worker's future behavior. In this paper, we propose a method for a machine learning agent to explain the target of the agent's own actions. The agent grasps the target of its own actions by predicting the result of the actions, and gives an appropriate expression for the predicted results by estimating the meanings of expressions on the basis of the shared reward.

1 はじめに

協調作業では、参加者が互いの行動目標を理解する必要がある。互いの行動目標を理解することで、行動の競合を回避し、協力や分担を選択することができるようになる。そのためエージェントと人との協調作業では、人がエージェントの行動決定モデル (制御モデル) の持つ行動目標を理解できるようにしなければならない。

しかし、深層強化学習によって獲得された制御モデルの行動目標を人が理解するのは難しい。深層強化学習によって獲得された制御モデルは大量のパラメータからなるネットワークで表現される。そのためネットワーク自体に着目して行動目標を理解することは容易でない。一方で制御モデルは、次時刻での行動という短期的な情報しか出力せず、またセンサ情報やモータへの入力といった低レベルで高次元な入出力情報を扱う。そのため人が制御モデルの入出力情報のみに着目して行動目標を理解するのも困難である。

*連絡先：慶應義塾大学理工学部情報工学科
〒 223–0061 神奈川県横浜市港北区日吉 3–14–1
E-mail: fukuchi@ailab.ics.keio.ac.jp

[1] は、エージェントの振る舞いの履歴から制御モデルの行動方策をマルコフ決定過程として抽出し、自然言語でエージェントの行動方策に関する質問を受け付け回答することに成功している。しかしこれは、エージェントの行動方策が固定されていることが前提として、環境の状態とエージェントの行動の対応を概略的に説明する方法である。そのため、例えば行動方策を動的に変化させるエージェントを対象に、協調作業の最中に行動目標を逐次説明する場合を扱うことはできない。また設計者が、説明表現を定義する 2 値分類器を用意する必要がある。

そこで本稿では、行動目標の説明表現を環境の中で動的に獲得し、各場面の中で自らの行動目標の説明表現を逐次出力する手法 (SETP+RWSG) を提案する。提案は、行動による環境遷移の予測から行動目標を把握する自己モデルの獲得 (Self-model with Environment Transition Prediction SETP) と、他者が与える説明表現と環境の報酬から行動目標に対する説明表現を生成 (Reward Weighted Symbol Grounding RWSG) の二段階からなる。SETP では、制御モデルの出力そのものではなく、エージェントの行動によって生じる環境の

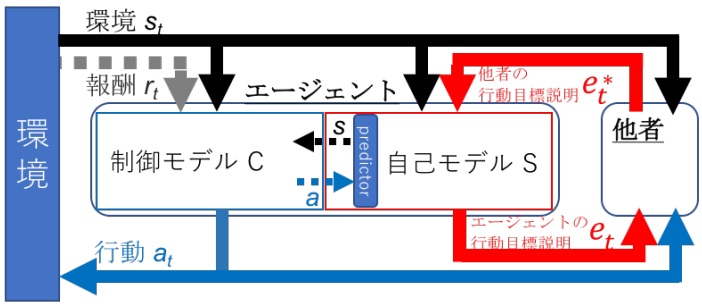

図 1: 設定の概要

遷移に着目することで、制御モデルの行動目標をメタ的に把握する。RWSG では、エージェントと他者の二者がいる状況を考える。他者はエージェントとは別に自身の行動目標を持ち、行動目標に対する説明を出力する。エージェントは、環境からの報酬が大きいほど、自らの持っていた行動目標が他者の説明に従っている可能性が高いという仮定をもとに、自らの行動目標と説明表現を対応関係を得る。そして得られた対応関係をもとに、SETP で把握した自らの行動目標に対する説明表現を決定する。

2 SETP+RWSG

2.1 設定

本稿で扱う設定の概要を図 1 に示す。設定では、エージェントと他者の二者がいる状況を考える。エージェントは環境 s_t を入力として行動 a_t を決定する制御モデル C と、制御モデルの行動目標を把握する自己モデル S からなる。エージェントは自己モデルが把握した制御モデルの行動目標に対し、説明表現 e_t を与え、出力する。e_t は数値を要素にもつベクトルの形で表現される。他者は、自らをエージェントの立場に置き換えた時の行動目標を持ち、行動目標に対する説明表現 e_t^* を出力する。エージェントは、e_t^* と、e_t^* が意味する行動目標の対応関係を推定することで、e_t^* をもとに自らの行動目標を説明できるようになる。

2.2 SETP

本稿では行動目標を、その後の行動による環境の変化 $\Delta s_{t:t+n} = s_{t+n} - s_t$ と定義する。SETP では、環境の変化の予測器 $Predictor$ を獲得して用いることで、エージェントの行動目標を把握する。

$$C(s_t) = a_t$$

$$Predictor(s_t, a_t) = \bar{s}_{t+1}$$

$$C(\bar{s}_{t+1}) = \bar{a}_{t+1}$$

$$Predictor(\bar{s}_{t+1}, \bar{a}_{t+1}) = \bar{s}_{t+2}$$

$$\vdots$$

$$C(\bar{s}_{t+n-1}) = \bar{a}_{t+n-1}$$

$$Predictor(\bar{s}_{t+n-1}, \bar{a}_{t+n-1}) = \bar{s}_{t+n}$$

$$\Delta\bar{s}_{t:t+n} = \bar{s}_{t+n} - s_t$$

2.3 RWSG

RWSG では、他者が出力する説明表現 e_t^* をもとに、エージェントの行動目標に対し説明表現を与える。まずエージェントの行動目標 $\Delta s_{t:t+n}$ と、その際の他者の行動目標の説明 e_t^* の組 $(\Delta s_{t:t+n}, e_t^*)$ を収集する。次に、収集した $\Delta s_{t:t+n}$ に着目し、クラスタリングを行う。そして得られたクラスタそれぞれに対して、行動目標の説明表現 e_t を与える。

ここで $\Delta s_{t:t+n}$ に対する環境からの報酬が大きいほど、$\Delta s_{t:t+n}$ が e_t^* に従っている可能性が高いと考える。エージェントと他者が同じ行動目標を持っている保証はないため、そのままでは $\Delta s_{t:t+n}$ と e_t^* を対応付けることはできない。しかし、両者は協調タスクの中で同じ報酬の最大化を目標としているため、$\Delta s_{t:t+n}$ に対する報酬が高い際は $\Delta s_{t:t+n}$ と e_t^* が対応関係にある可能性が高いと期待される。そこで、まずそれぞれのクラスタに含まれる $(\Delta s_{t:t+n}, e_t^*)$ の中からエピソード内の報酬が大きかったものを抽出する。そして抽出した組の中で、他者からの戦略が与えられた期待値を計算する。最後に得られた期待値をクラスタ間で正規化を行い、エージェントによる行動目標の説明 e_t とする。

3 実験

SETP+RWSG によって出力されるエージェントの行動目標の説明 e_t を検証した。実験環境には、Open AI gym[2] が提供するゲーム Lunar-Lander v2 を利用し、Deep Q Network[3] によってクリア率が 9 割程度になるまで学習したエージェントを対象に SETP+RWSG を行った。Lunar-Lander v2 はロケットを月面に軟着陸させるゲームで、エージェントは「右・下・左へのジェット噴射/何もしない」の 4 通りの操作を毎フレーム選択することでクリアを目指す。他者の行動目標の説明 e_t^* は、「落下速度を減少/そのまま/増加」と「左/そのまま/右へ移動」の 2 種類 9 通りで (図 2)、人手によりあらかじめ定めた。

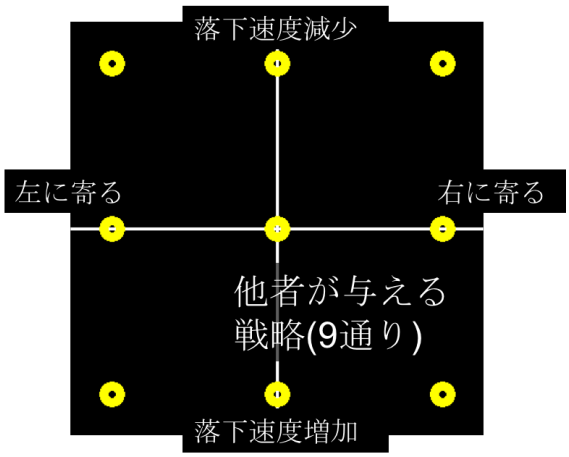

図 2: 他者が与える行動目標の説明

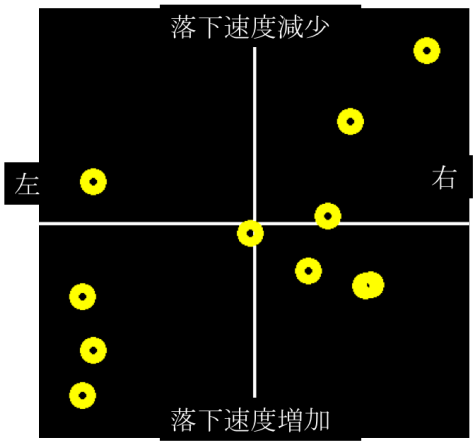

図 3: RWSG によって獲得された説明表現

3.1 結果

RWSG の結果、各クラスタを説明する表現は図 3 のようになった。他者の戦略は左右対称に定められているにも関わらず、獲得された表現は左右非対称になった。エージェントの振る舞いを観察すると、エージェントは右回りする際よりも左回りする際のほうが早く落下することがわかった。図 3 の結果は、エージェントが持っている行動特性に合わせて、行動目標の説明表現が動的に獲得できていることを示していると考える。

また、図 4 は、あるエピソードでのエージェントの振る舞いと、その際に SETP+RWSG によって出力されたエージェントの行動目標、他者が与えていた行動目標をそれぞれ可視化したものである。このエピソードでは、エージェントに右への初速度が大きく与えられていて、エージェントは大きく右に振られながら月面に着陸している。

他者の与える説明は、「左へ向かう」「速度を減少させる」を意味する左上に固まっている。一方でエージェントの説明は、自らの行動目標に合わせて適宜出力を変えることができていることがわかる。

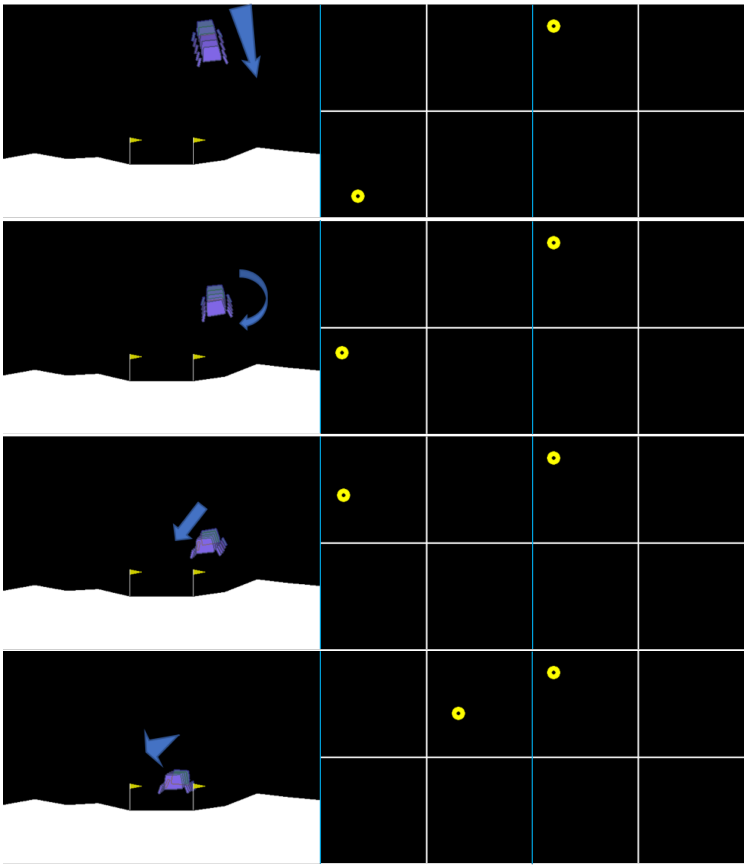

図 4: あるエピソードでのエージェントの振る舞い (左)、SETP+RWSG によって得られたエージェントの行動目標 (中)、他者が与えていた行動目標 (右)

4 終わりに

本稿では、行動による環境遷移の予測から行動目標を把握する自己モデルの獲得 (SETP) と、他者が与える説明表現と環境の報酬から行動目標に対する説明表現を生成 (RWSG) の二段階によって、行動目標の説明表現を環境の中で動的に獲得し、各場面の中で自らの行動目標の説明表現を逐次出力する手法を提案した。

実験ではエピソードクリア率 9 割程度の、ある程度行動獲得が済んだエージェントに対して手法を適用した。現在は、RWSG はより未熟なエージェントにも適

用できるのではないかと考え、検証を予定している。より未熟な段階で行動目標と説明表現の対応関係が得ることができれば、他者の行動目標の説明をもとに、運動空間を探索する際のバイアスとしても利用できると考えている。

参考文献

[1] Bradley Hayes and Julie A. Shah: Improving Robot Controller Transparency Through Autonomous Policy Explanation, In *Proceedings of the 12th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 303-312 (2017).

[2] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba.: Openai gym. *arXiv:1606.01540* (2016).

[3] V.Mnih, K.Kavukcuoglu, D.Silver, A.Rusu, J.Veness, et al.: Humanlevel control through deep reinforcement learning ”, *Nature*, 518, pp. 529-533 (2017).