人工知能学会研究会資料
SIG-AGI-006-07

異種エージェントへの教示に向けたInstruction-based Behavior Explanationの応用の検討

A Study on Application of Instruction-based Behavior Explanation for Teaching to Heterogeneous Agents

福地庸介[1, *] 大澤正彦[1, 2], 山川宏[3, 4], 今井倫太[1]
Yosuke Fukuchi[1], Masahiko Osawa[1, 2], Hiroshi Yamakawa[3, 4], Michita Imai[1]

[1] 慶應義塾大学
[1] Keio University
[2] 日本学術振興会 特別研究員 (DC1)
[2] Research Fellow of Japan Society for the Promotion of Science (DC1)
[3] 株式会社ドワンゴ　ドワンゴ人工知能研究所
[3] DWANGO Co., ltd Dwango Artificial Intelligence Laboratotry
[4] 全脳アーキテクチャ・イニシアティブ
[4] The Whole Brain Architecture Initiative

Abstract: Under large state and action spaces, it is difficult for a reinforcement learning agent to learn the agent's policy within a practical time. Previous studies have proposed methods in which a trainer gives better actions to a trainee to promote the learning. However, when action spaces of a trainer and a trainee is not the same, the instruction does not work without mapping from the instruction to the trainee's variable space. In this paper, we deal with three types of instruction: action-based expression, abstract expression from a human trainer, and expression output by Instruction-based Behavior Explanation, which is a framework to announce a reinforcement learning agent's future behavior. The three instructions were mapped to agents' action spaces with deep reinforcement learning, and we compared the mappings to consider the form of information towards heterogeneous agents' instruction.

1　はじめに

深層強化学習は，多くのタスクでエージェントの自律的な行動学習を実現し，優れた成績を挙げてきた [1].

一方強化学習は，複雑な状態･行動空間の下では，行動学習を現実的な時間内で行うことが困難になる [2].

学習時間の問題に対して，学習中のエージェントに教示を与えることで学習を促進する方法が提案されてきた．教示者は人間とエージェントの両者が考えられ，教示の形は大きく 4 種類ある．模倣学習 [3] は，教示者が与える行動軌跡のサンプルを学習に利用する方法で，教示は強化学習における状態の形で与えられていると言える．Reward Shaping[4] は，教示者がエージェントの振る舞いに報酬の形で追加のフィードバックを与える方法である．Policy Shaping[5] は，行動探索時のエージェントの行動決定を教示者がより良い行動で上書きすることで，有望な行動探索を促す方法である．教示の情報は行動の形で与えられる．さらに，自然言語など，強化学習が扱うシグナルと異なる抽象的な表現による教示が提案されている [6]．抽象表現による教示では，エージェントに固有な行動や報酬について考慮する必要がないため，エージェントの設計を知らない人間からの教示でも自然に実現できる．

しかし Reward Shaping では，教示として与えられる追加報酬の不定期性や一貫性の欠如が学習を阻害することあり [7]，人間からの教示を扱う際特に問題とな

*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

る．一方行動をベースとした教示は，行動空間を共有していないエージェント (異種エージェントと呼ぶ) に対して直接適用することはできない．教示者が与える教示を，学習者が扱う変数空間に接地する必要が生じるためである．教示表現の接地が必要なのは，強化学習の枠組外の表現である抽象表現による教示でも同様である．

そこで本稿ではまず，異種エージェントからの行動による教示，人間が設定した抽象表現からの教示，そして人から与えられた抽象表現による教示をもとに強化学習エージェントの振る舞いを予告する枠組 Instruction-based Behavior Explanation (IBE)[8] の出力による教示の 3 種類の教示を用意した．そして教示表現からエージェントの行動空間へのマッピングを深層強化学習を用いて行い，それぞれの教示でのマッピングの結果を比較することで，異種エージェントへの教示に適した情報の形を考える．人–機械学習エージェント間コミュニケーションも一種の異種エージェント間コミュニケーションである．人–エージェント間の適切な情報のやり取りによる相互理解の実現は，人とエージェントの協調作業や，エージェントの意図せぬ振る舞いによる事故の防止にも通じる．

2 背景

2.1 強化学習

本稿では，強化学習エージェントを扱う．エージェントは，各時刻環境の状態 s_t を受け取り，方策 π を元に行動空間 A から行動 a_t を選択する．$\pi(s_t, a)$ は，エージェントが状態 s_t において行動 a を選択する確率を表す．環境はエージェントの振る舞いに対し報酬 r_t を与える．強化学習におけるエージェントの目標は，割引率を γ として，将来得られる累積報酬 $R = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$ を最大化する最適方策を獲得することである．

2.2 Instruction-based Behavior Explanation (IBE)

IBE は，強化学習エージェントの将来の振る舞いを予告するための枠組である．IBE では，エージェントの行動学習の際に与えられた抽象表現による教示を再利用することで，エージェントの振る舞いを予告する．

IBE による振る舞い予告の流れは 3 段階に分かれる．まずエージェントの行動による環境の変化 $\Delta s_t = (s_{t+n}, s_t)$ から教示に用いられている表現空間 m へのマッピング f を獲得する．さらに Δs_t をエージェントの方策 π と環境の変化の予測器 $predictor$ を用いた内的シミュレーションによって推定する．そして得られた f と Δs_t から計算される $f(\Delta s_t)$ を，エージェントの振る舞い予告表現とする．

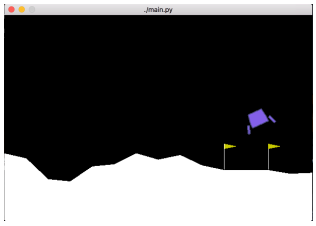

図 1: ゲーム環境

3 検証する教示

本稿では教示表現からエージェントの行動空間へのマッピングを深層強化学習を用いて行い，得られたマッピングを比較することで，異種エージェントへの教示により適した情報の形を考える．

行動による教示は教示者と学習者の行動空間が一致している場合は，教示と対応する行動がマッピングされることで教示の持つ意味を完全に共有できる．一方教示者と学習者の行動空間が一致していない場合，教示と行動の対応性が下がるため，マッピングは難しくなると考えられる．また教示者の行動に対応する学習者の行動が存在しない場合，本来教示が持つ意味と異なる行動に教示が解釈されることになる．そのため行動空間の違いが大きいほど，行動による教示は不利になると推測される．

抽象表現による教示は，エージェントに固有の行動空間に依存しない表現による教示が可能である．そのため行動空間の異なるエージェントへの教示の際は，行動による教示よりも効果的だと期待される．

学習者の状態を教示者に与え IBE を利用すれば，教示者が学習者の立場に立った際の振る舞いの予告を得ることができる．そこで本稿では，IBE を教示者に適用し，出力される予告を教示として利用する．IBE による教示により，人手で作り込んだ教示よりも柔軟に各場面での教示を出力できると期待される．

4 マッピングによる検証

4.1 実験設定

ゲーム Lunar-Lander v2 をベースとしたゲーム環境を用意した (図 1)．Lunar-Lander は宇宙船エージェントを月面の着地点に軟着陸させることを目指すゲームである．ゲーム内で宇宙船は，毎フレーム以下の 4 通りの操作を選択することで，動きを制御する．

操作: {0: 何もしない, 1: 左へのジェット噴射,
2: 下へのジェット噴射, 3: 右へのジェット噴射 }

検証のため，行動空間が異なるエージェントタイプ A,B を用意した．タイプ A は，行動空間 A_A から 2 フ

レーム分の操作を 2 フレームごとに選択する．

$$A_A = \{(0,0),(1,2),(2,2),(3,2)\}$$

タイプ B は，行動空間 A_B から 4 フレーム分の操作を 4 フレームごとに選択する．

$$\begin{aligned} A_B = \{&(0,0,0,0),(0,0,1,1),(0,0,2,2),(0,0,3,3),\\ &(1,1,0,0),(1,1,1,1),(1,1,2,2),(1,1,3,3),\\ &(2,2,0,0),(2,2,1,1),(2,3,2,2),(2,2,3,3),\\ &(3,3,0,0),(3,3,1,1),(2,3,2,2),(3,3,3,3)\} \end{aligned}$$

行動による教示を用意するため，タイプ A のエージェント 5 体を教示なしで学習させ，最もスコアが高いモデル (教示モデルと呼ぶ) を得た．

抽象表現による教示は，教示モデルによるエージェントの振る舞いの観察から著者らが決定した．抽象表現による教示は，エージェントの落下速度を v_y，着地点からの高さを y，傾きを θ として以下の式から得られる値をタプルとして出力した．

$$m_t^1 = \begin{cases} -1 & \text{if } v_y < 0.8 \cdot (y-0.7)^2 - 0.55 \\ +1 & \text{else if } v_y > 1.1 \cdot (y-0.8)^2 - 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$m_t^2 = \begin{cases} -1 & \text{if the agent is right of the right flag} \\ +1 & \text{else if the agent is left of the left flag} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$m_t^3 = \begin{cases} -1 & \text{if } \theta > 0.12 \cdot y + 0.05 \\ +1 & \text{else if } \theta < -0.12 \cdot y - 0.05 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$m_t^4 = \begin{cases} +1 & \text{if the agent has landed} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

各教示はそれぞれ「落下スピードを下げる/上げる/そのまま」「左へ向かう/右へ向かう/そのまま」「時計回りに回転する/反時計回りに回転する/そのまま」「静止する/指示なし」を意味する．

更に教示モデルに IBE を適用し，教示モデルが学習者の状況にある際の教示モデルの将来の振る舞いを予告できるようにした．IBE のマッピング f を獲得する際に利用した教示は，式 1–4 を利用した．そして IBE による振る舞いの予告を教示として，エージェントに学習させた．

マッピングは，教示を入力として各エージェントタイプの行動空間の行動を出力とする深層強化学習によって獲得させた．マッピングは学習 100 エピソードごとに，30 エピソードのテストの中で検証し，1 エピソード内で得られる報酬の合計をスコアとして比較した．

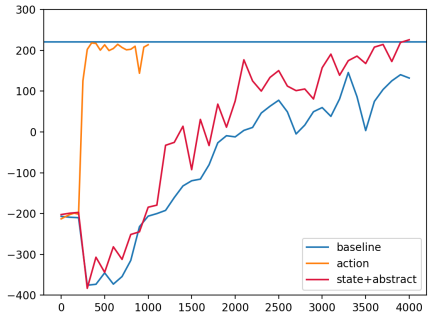

図 2: 学習により経過したエピソード数 (横軸) と，タイプ A のスコアの平均 (縦軸)．横線は教示モデルのスコア

4.2 結果 1 同種エージェントへの教示

教示モデルと同じ行動空間 A_A を持つタイプ A のエージェントの入力に，状態のみのベースライン (教示なし)，教示モデルの行動による教示，状態と人間が決定した教示を与えた際のスコアを図 2 に示す．教示者の教示空間と訓練者の行動空間は完全に対応しているため，行動による教示とエージェントの行動はすぐに結び付けられ，教示モデルと同じスコアを示すようになった．人間が与えた教示も，教示なしと比べて一定の学習促進効果が見られた．しかし行動による教示とは大きく差がついた．行動空間を共有する同種のエージェントへの教示では，行動による教示が効果的だと考えられる．

4.3 結果 2 異種エージェントへの教示

教示モデルと行動空間が異なるタイプ B のエージェントに行動による教示を与えた場合，行動による教示と状態を与えた場合で，マッピングを行った．結果を図 3 に示す．入力に行動のみを与えた際は，学習中盤の長い間でスコアが伸び悩んだ．教示者の教示空間と訓練者の行動空間が対応関係が乏しさが，マッピングを難しくしていることがわかる．入力に行動と状態の両方を与えた場合は，学習中盤まで教示の効果は見られなかった．一方学習後半では，スコアの伸びが停滞するベースラインに対しての優位性がみられた．結果から，教示空間と行動空間の対応関係の乏しさにより，教示空間から行動空間へのマッピングに時間がかかったと考えられる．教示空間と行動空間の対応関係がより大きくなれば，マッピングにかかる時間はさらに悪化する可能性がある．

次に，人間が決定した抽象表現による教示と IBE による教示を与えた際の結果を図 4 に示す．人間が決定

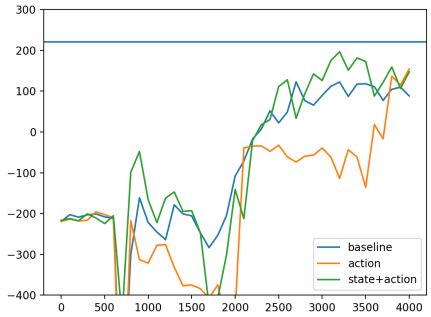

図 3: 行動による教示を与えた際のタイプ B のスコアの平均.

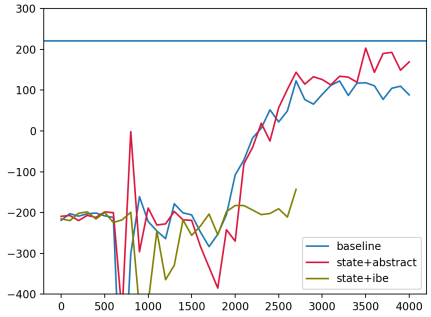

図 4: 抽象表現による教示を与えた際のタイプ B のスコアの平均.

した教示について，学習中盤はベースラインと同様のスコアで，学習後半に優位性が見られた．

しかし IBE の出力を教示とした場合はむしろ 2,500 エピソードまでスコアが伸びなかった．教示が機能しなかった理由の一つに，体勢が大きく崩れてしまうと教示モデルでも復帰が困難になるため，教示モデルが振る舞いを予告しても有効な教示にならなくなることが考えられる．対策として，IBE の *predictor* によって教示者が訓練者の立場に立った際の振る舞いを予測する際，教示者が現状から復帰できるかも判断できるので，復帰困難の際はエピソードを取りやめ，次のエピソードに進んでしまうという方法が考えられる．

5　おわりに

行動による教示，人間が決定した抽象表現による教示，IBE による教示の 3 種類を用意し，深層強化学習によって，教示表現から学習者の行動空間へのマッピングの獲得を行なった．そしてそれぞれの教示でのマッピングの結果を比較することで，異種エージェント間での教示に適した情報の形を考えた．

結果，行動空間の異なるエージェントへの教示における抽象的な表現による教示の有効性が示唆された一方で，同じ抽象表現による教示である IBE による教示は機能しなかった．今後は IBE による教示を改良し，より良いエージェントへの教示の実現を目指したい．

参考文献

[1] Mnih, V., et al. Kavukcuoglu, K., Silver, D., et al.: Humanlevel control through deep reinforcement learning. *Nature*, 518, pp. 529–533, 2017.

[2] Knox, W. B., Stone, P.,: Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the fifth international conference on Knowledge capture*, ACM, pp. 916, 2009.

[3] Ho, Jonathan, and Stefano Ermon: Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, pp. 4565–4573, 2016.

[4] Pilarski, P. M., Dawson, M. R., Degris, T., et al.: Online human training of a myoelectric prosthesis controller via actor-critic reinforcement learning. In *Proceedings of the IEEE ICORR*, pp. 1-7, 2011.

[5] Griffith, S., Subramanian, K., Scholz, J., et al.: Policy Shaping: Integrating Human Feedback with Reinforcement Learning. In *Advances in Neural Information Processing Systems 26* pp. 2625–2633, 2013.

[6] Peng, B., MacGlashan, J., Loftin, R., et al.: A need for speed: Adapting agent action speed to improve task learning from non-expert humans. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems* pp. 957–965 2016.

[7] Ng, A. Y., Harada, D., Russell, S.: Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the 16th ICML*, pp. 341-348, 1999.

[8] Fukuchi, Y., Osawa, M., Yamakawa, H., et al.: Application of Instruction-based Behavior Explanation to a Reinforcement Learning Agent with Changing Policy In *Proceedings of the 24th International Conference on Neural Information Processing*, 2017