Vol.2020-ICS-200 No.5
2020/9/14

深層強化学習を用いたサッカータスクにおける行動獲得に関する考察

阿部 宇志[1,a)]　折原 良平[1]　清 雄一[1]　田原 康之[1]　大須賀 昭彦[1]

概要：これまで, 深層学習と強化学習を組み合わせた深層強化学習がチェスや囲碁, 将棋といったゲームに対して適用され大きな成果が見られてきた. それに加え, 近年ではサッカーゲームのような複数のエージェントが関わる問題に対して強化学習を行う研究が取り上げられてきている. サッカータスクのようなマルチエージェント強化学習は, 自らの行動以外に加味する情報が多くあり, 学習が難しいタスクであることから研究が盛んに行われている. 本研究ではその中でも, DeepMind 社の物理演算エンジン Mujoco のサッカータスクを用いた深層強化学習での行動獲得を目指す. サッカータスクは報酬がスパースであるタスクであり, 外部報酬だけでは学習に膨大な時間がかかるため, 様々な研究がされてきた. ここでは, 複雑なタスクに対応するために, スパースである外部報酬だけでなく, 段階的に報酬を与える Shaping 強化学習を行うことで, 学習に及ぼす影響を考察する.

1. はじめに

1.1 背景

これまで, 深層学習と強化学習を組み合わせた深層強化学習が, ゲーム AI などで大きな成果をもたらしている. それに加え, 近年ではサッカーゲームのような複数のエージェントが関わる問題に対して強化学習を行う研究がなされ, サッカーを取り上げた研究も盛んに進められている. サッカーは一人の実力だけで勝利することは極めて困難であり, パス交換や組織的な守備などにおける協調行動が必要であることから, マルチエージェント強化学習の題材として取り上げられることが多い.

最も知られているサッカーシステムの RoboCup サッカーはこれまで 2D シミュレーション [1] だけでなく, 実際にロボットを使ってより人間に近い環境で研究が行われている. また, DeepMind 社によって作成された物理演算エンジン Mujoco[2] を使ったサッカータスクや, Google 社が作成した Google Research Football Environment[3] など, サッカーを扱うシミュレーションが多く存在し, これらのシミュレーターをもとに研究が活発に行われてきた.

サッカータスクのようなマルチエージェント強化学習は, 自らの状態だけでなく, 味方や敵の行動や状態を加味して, 敵対行動と協調行動をとる必要があるため, 学習が難しいタスクであることから研究が数多く行われてきた. その中でもサッカータスクは報酬がスパースな環境として知られている. 報酬がスパースな場合, 報酬を得るまでの探索に膨大な学習時間を要し, 学習が困難になるため [4], 様々な工夫が取り入れられている. 外部報酬だけでなく, 予測した環境と実際の環境の差分を利用する内発的動機付け [5][6] を導入することや, 人間の経験を学習に取り入れるヒューリスティクスを導入することで学習の効率化が可能であることが確認されてきた.

本研究ではそういった学習の工夫の中でも Shaping 強化学習に注目し, 深層強化学習を用いたエージェントの行動獲得を行う. Shaping 強化学習は, 本来のゴールにたどり着くまでに階層的に報酬を与えることで, 学習の効率化を目指す手法である [7]. 報酬がスパースに与えられるタスクでは, ゴールにたどり着くまでに報酬が与えられる機会が少なく, 探索に膨大な時間がかかる. そこでサッカーのような複雑で探索に時間がかかるタスクに対しては, 段階的に報酬を与え, 効率よく学習が行われる研究がされてきた.

Shaping 強化学習では, 適切に報酬設計をすることができれば学習を効率化できるが, 不適切な報酬設計をすると, 学習が進まなくなってしまう懸念がある. そのため, Rewawd Shaping を行う場合, 設計者の豊富なドメイン知識を使った適切な報酬設計が必要になる. 本研究では, 物理演算エンジン Mujoco のサッカータスクで, Shaping 強化学習を深層強化学習に取り入れて実験をし, Reward Shaping がサッカーという複雑なタスクでの学習に対してどのような影響

[1] 電気通信大学 大学院情報理工学研究科
Graduate School of Informatics and Engineering, The University of Electro-Communications
[a)] abe.takashi@ohsuga.lab.uec.ac.jp

を及ぼすのか, 考察を行った.

1.2 論文の構成

本論文の構成は次の通りである. 2 章で関連研究について述べる. 3 章では強化学習に関して, 4 章では提案手法について述べる. 5 章で評価について, 6 章で本論文のまとめを記す.

2. 関連研究

2.1 サッカーの強化学習

マルチエージェント強化学習の代表例として, サッカーにおける強化学習の研究はこれまで多くなされてきた.

数あるサッカータスクの中でもよく使われている RoboCup サッカーのうち, 特に Stone らによって行われた Keepaway タスクの研究 [8] は, これまでの RoboCup サッカーの戦略構築で多く用いられてきた. Keepaway タスクは, ボールを保持する“the keepers”チームとボールを奪い取る“the takers”チームによって構成されるタスクで,“the takers”チームにボールを奪われないように“the keepers”チームがパスやドリブルによってボールを保持する行動を学習するタスクである. 現実のサッカーである状況を想定して練習を行うように, エージェントの行動獲得においても, Keepaway タスクのように, 状況が限定されたタスクで事前学習を行うことが必要となっている.

また, DeepMind 社によって作成された物理演算エンジン Mujoco のサッカータスクでも研究がなされてきた. Liu らによる研究 [9] でのタスクでは, 攻撃側と防御側プレイヤーがそれぞれ 2 体の 2 対 2 で攻撃と守備を学習している. 現実の試合でも 2 対 2 という局面が攻撃守備の両面で頻繁に見られ, 攻撃に関してはパスやボールを受ける動き, 守備に関してはチャレンジ＆カバー（複数人で守備をする場合に, 相手選手にプレッシャーをかける選手とそのカバーを行う選手の役割分担を行うこと）をエージェント自身が学習するために, 有用なタスクであるといえる.

その他にも, Mujoco のサッカータスクでの内発的動機付けの効果が Chitnis らの研究 [10] によって実証されている. 報酬がスパースなタスクに対して, 外部報酬だけでなく, 予測した状態との違いを“Surprise”とし, 内発的報酬として外部報酬に加えて学習することで学習の効率化に成功した. 本研究では, これらと同様に物理演算エンジン Mujoco のサッカータスクを用いて, 考察を行った.

2.2 Shaping 強化学習

これまで, 設計者によって設計された報酬が段階的に与えられる, Reward Shaping を用いた Shaping 強化学習の研究 [7] が行われてきた.

サッカーの強化学習においても, Shaping 強化学習を用いる研究がこれまで多くされてきた. RoboCup では, Keepaway タスクに対して Reward Shaping を用いることで, 学習が効果的に行われることが Devlin らの研究 [11] によって実証されている. この研究では, Taker 同士の距離を報酬として与えること, 守備の役割を促すように報酬を与えることで, 守備の向上が見られた.

Shaping 強化学習においては, 不適切な報酬設計をすると学習を悪化させる原因となってしまうため, 設計者のドメイン知識による適切な報酬設計が必要である. 本研究では, サッカーという学習が困難なタスクを深層強化学習で学習するために, Reward Shaping を用いて協調行動を行い, 学習に及ぼす影響を考察する.

3. 強化学習

3.1 強化学習手法

強化学習ではモデルフリー, つまり環境の知識がない状態から学習するため, 探索を行いながら状態の特徴を適切に抽出することが求められる. また, 特徴を適切に把握するだけでなく, その後の状態を先読みして行動選択する必要がある. これらの重要な役割を深層学習が担い, これまで制御が難しいとされていたタスクに対して, 深層学習を強化学習に適用した学習方法が, 深層強化学習として知られている.

本研究で取り扱うタスクは, 行動を表している変数が連続変数であり, 強化学習の中でも広く知られている Q 学習 [12] を適用するのは難しい. このような連続値の出力が求められるタスクの場合, 行動価値を推定するのではなく, 行動の確率分布を表す方策を学習することが有効である. そのため, ここでは DQN[13] のような価値ベースの強化学習手法ではなく, 方策勾配法のような方策ベースの強化学習手法で研究を行う.

3.2 方策勾配法

エージェントの行動を最適化するために, 方策反復法で最適方策を見つける方法がある. 方策反復法は, ある方策のもとで価値関数を計算する方策評価ステップと, そこで得た価値関数を最大化するように方策を更新する方策改善ステップを繰り返して最適方策を見つける手法である [14].

モデルフリーな方策反復法に対してのアプローチとして, 方策勾配法があげられる. 方策勾配法は, 目的関数 $J(\theta)$ における方策パラメータ θ の勾配 $\nabla_\theta J(\theta)$ の方向に θ を更新して, 目的関数 $J(\theta)$ がより大きな値をとるように改善する方策ベースの手法である. 勾配 $\nabla_\theta J(\theta)$ は以下のように示される [15].

$$\nabla_\theta J(\theta) = \mathbb{E}_\pi[(\nabla_\theta \log \pi(A_t|S_t,\theta))Q_\pi(S_t,A_t)] \quad (1)$$

ここではある時間 t で, 状態 S_t における行動 A_t を選択したときの価値関数を $Q_\pi(S_t,A_t)$ としている. これにより, Q 関数による方策評価を取り込んだ方策改善をすることがで

きるようになっている.

3.3 REINFORCE アルゴリズム

本研究では, REINFORCE アルゴリズム [16] を用いている. REINFORCE アルゴリズムは, 方策勾配法において, 行動価値関数を割引報酬和 G_t で近似することで学習を行うアルゴリズムである. ある時間 t における状態 S_t から, 行動 A_t を選択することで将来的に得られる報酬 R_{t+k} とし, 割引率を γ とすると, 割引報酬和 G_t, 方策パラメータ θ の勾配 $\nabla_\theta J(\theta)$ は以下のように表される.

$$G_t = \sum_{k=1}^{T-t} \gamma^{k-1} R_{t+k} \tag{2}$$

$$\nabla_\theta J(\theta) \approx \frac{1}{T} \sum_{t=0}^{T-1} \nabla_\theta \log \pi(A_t|S_t) G_t \tag{3}$$

REINFORCE アルゴリズムでは, ある時間 t における行動価値関数を将来的に得られる報酬 R_{t+n} の割引率を考慮した総和である割引報酬和 G_t に置き換えている.

4. 提案手法

マルチエージェントタスクであるサッカータスクに対して, 方策勾配法である REINFORCE アルゴリズムを用いて, Shaping 強化学習を行った.

4.1 想定環境

マルチエージェントの協調行動による学習の影響を考察するため, 実験は Chitnis らによって Mujoco のサッカータスクで行われた研究 [10] に似せた想定環境で行った. **図 1** に想定環境の図を示す. ここでは, 2 体のエージェントの相互作用を確認するため, 中央に置かれたボールを 2 体のエージェントが関わってゴールを決めた時に目標達成とする. そのため, 片方のエージェントがボールに触った後, もう片方のエージェントがボールに触れてゴールを決めなければ, 単独のエージェントだけでゴールを決めたとしてもゴールと見なされない. ボールの受け渡しというエージェントの相互作用が必須となっている.

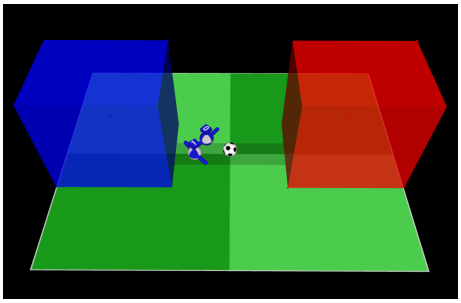

図 1: 本実験での想定環境

4.2 Reward Shaping

今回の環境では, 以下のように報酬を与えることとする.

- 2 体のエージェントが触れてゴール +10
- 1 体目のエージェントが初めてボールに触れる +1
- 2 体目のエージェントが初めてボールに触れる +1

今回の環境の目標は 2 体のエージェントが触れてゴールすることであるため, 1 体しか触れずにゴールに入った場合は報酬を与えないこととしている. Reward Shaping として, 各エージェントがそのエピソードのうち初めてボールに触った時に報酬を与える. Shaping の報酬は小さめに, 最終目標の報酬は大きめにすることでボールに触れるだけで学習を収束させず, 最終目標まで向かわせている.

Shaping 強化学習では, 適切な報酬設計が必要になる. 本研究での想定環境では, ボールの受け渡しをするというシンプルな行動の学習がねらいであるため, 両エージェントがボールに触れるように Reward Shaping を行い, パスに近い協調行動を学習できるように促した.

5. 評価実験

5.1 実験設定

本研究では, DeepMind 社で作成された物理演算エンジン Mujoco のサッカータスクを用いて実験を行う. 4.1 で述べたように, 2 体のエージェントの協調行動を見るため, 2 体のエージェントがボールに触れてからゴールすることで目標達成と見なし, 報酬を与えることとする.

実験では, 味方の 2 体のエージェントのみで行動獲得を行うこととする. 状態は味方やボール, ピッチの見え方等を示す 81 次元の数値で表され, 行動は 3 次元の数値で決定される. エピソードは 30 秒を 1 エピソードとし, ゴールに入った時にエピソードは終了する.

モデルはガウスモデルによる確率的方策と価値関数の 2 つのモデルを用意し, 状態を認識したのちに確率的方策モデルから行動を獲得, その行動によって得られた報酬と価値関数モデルから得られる価値と比較して行動の評価を行い, モデルの更新を行う. 方策モデルは入力が 81 の状態, 隠れ層は 3（それぞれ 810, 220, 60 ノード）, 出力は 3 次元の行動とする. 価値関数モデルは入力が 81 の状態, 隠れ層は 3（それぞれ 810, 63, 5 ノード）, 出力は 1 次元の行動の価値とする. 学習率はそれぞれ 0.99, 最適化関数は方策モデルが RMSProp, 価値関数モデルが Adam を使用している.

5.2 実験結果

5.2.1 平均目標達成率

図 2 に平均目標達成率を示す. ここでは学習初期である 10000 エピソードまでの実験結果を見て, 学習に関する考察を行った. 今回は Reward Shaping 無しの No Shaping と Reward Shaping 有りの Shaping の 2 パターンで比較実験を行った.

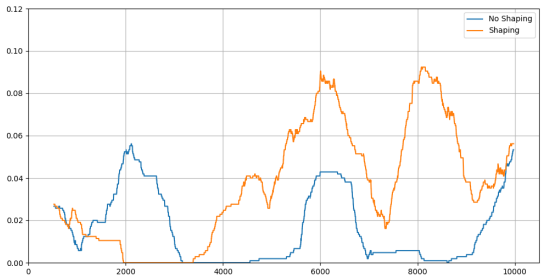

図 2: 平均目標達成率

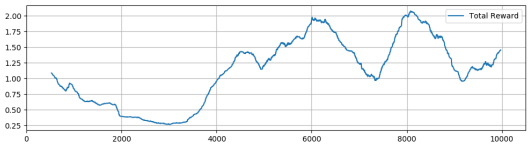

図 3: Shaping ありの平均総獲得報酬

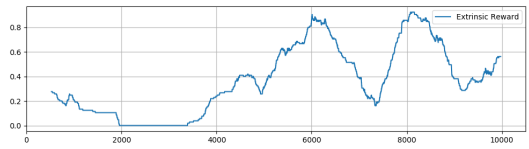

図 4: Shaping ありの平均外部獲得報酬

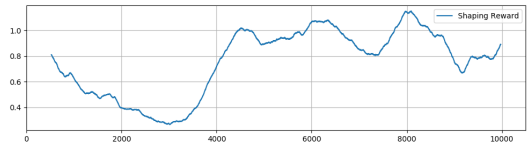

図 5: Shaping ありの平均 Shaping 獲得報酬

図 2 より, Shaping 有りの場合, 6000〜8000 エピソードまで学習すると, 目標達成率が 0.1 まで近づいていることがわかり, 3000 エピソード以降では Shaping 無しの場合に比べて, 目標達成率が高くなっている. 10000 エピソード全体を通しての目標達成は Shaping 有りが 403 回, Shaping 無しが 203 回で約 2 倍の達成率になっている. 学習初期であるため, 学習が不安定であるものの, Reward Shaping を行うことによって, 報酬獲得の回数が多くなることから, 学習初期の目標達成率も高くなることがわかる.

5.2.2 Shaping 強化学習による獲得報酬

図 3〜図 5 で Shaping ありの平均総獲得報酬, 外部獲得報酬, Shaping 獲得報酬を示す. 外部獲得報酬が 6000 エピソードで増加した後, 7000 エピソードで減少していることがわかる. しかし, Shaping 報酬は安定して獲得できており, その後の 8000 エピソードで再び外部獲得報酬の増加を促すことができていると言える.

また, 図 5 を見ると, 4000 エピソード以降では Shaping 報酬をおよそ 1.0 獲得できている. このことから, どちらか片方のエージェントは, 毎エピソードでボールに触れることができるように学習出来ていることがわかる. この Reward Shaping によって 4000 エピソード周辺から外部報酬を得られるようになってきており, 適切な報酬設計を組むことができているとわかる.

6. まとめ

本研究では, 物理演算エンジン Mujoco のサッカータスクで, 深層強化学習に Reward Shaping を取り入れることによる学習効果の影響を考察した. 学習初期段階での目標達成率を比較し, Reward Shaping を導入することで, 学習の改善を見ることができた. 今後は学習初期段階だけでなく, エピソード数を増やしての比較実験を行うことや, 内発的報酬動機付けの導入などマルチエージェント強化学習に対する工夫を加え, より効率的な学習を目指す.

謝辞 本研究は JSPS 科研費 JP17H04705, JP18H03229, JP18H03340, JP18K19835, JP19H04113, JP19K12107 の助成を受けたものです. 本研究を遂行するにあたり, 研究の機会と議論・研鑽の場を提供して頂き, 御指導頂いた早稲田大学 本位田真一教授, 鄭顕志准教授をはじめ, 活発な議論と貴重な御意見を頂いた研究グループの皆様に感謝致します.

参考文献

[1] Kitano, H., Asada, M., Kuniyoshi, Y., Noda, I. and Osawa, E.: RoboCup: The Robot World Cup Initiative, *Proceedings of the First International Conference on Autonomous Agents*, AGENTS '97, New York, NY, USA, Association for Computing Machinery, p. 340–347 (online), DOI: 10.1145/267658.267738 (1997).

[2] Todorov, E., Erez, T. and Tassa, Y.: MuJoCo: A physics engine for model-based control, *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033 (2012).

[3] Kurach, K., Raichuk, A., Stanczyk, P., Zajac, M., Bachem, O., Espeholt, L., Riquelme, C., Vincent, D., Michalski, M., Bousquet, O. and Gelly, S.: Google Research Football: A Novel Reinforcement Learning Environment, *CoRR*, Vol. abs/1907.11180 (online), available from ⟨http://arxiv.org/abs/1907.11180⟩ (2019).

[4] Sutton, R. S. and Barto, A. G.: *Introduction to Reinforcement Learning*, MIT Press, Cambridge, MA, USA, 1st edition (1998).

[5] Pathak, D., Agrawal, P., Efros, A. A. and Darrell, T.: Curiosity-Driven Exploration by Self-Supervised Prediction, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (2017).

[6] Agrawal, P., Nair, A. V., Abbeel, P., Malik, J. and Levine, S.: Learning to Poke by Poking: Experiential Learning of Intuitive Physics, *Advances in Neural Information Processing Systems 29* (Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I. and Garnett, R., eds.), Curran Associates, Inc., pp. 5074–5082 (online), available from ⟨http://papers.nips.cc/paper/6113-learning-to-poke-by-poking-experiential-learning-of-

intuitive-physics.pdf⟩ (2016).
[7] Ng, A. Y., Harada, D. and Russell, S. J.: Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping, *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., p. 278–287 (1999).
[8] Stone, P., Sutton, R. S. and Kuhlmann, G.: Reinforcement Learning for RoboCup Soccer Keepaway, *Adaptive Behavior*, Vol. 13, No. 3, pp. 165–188 (online), DOI: 10.1177/105971230501300301 (2005).
[9] Liu, S., Lever, G., Merel, J., Tunyasuvunakool, S., Heess, N. and Graepel, T.: Emergent Coordination Through Competition, *CoRR*, Vol. abs/1902.07151 (online), available from ⟨http://arxiv.org/abs/1902.07151⟩ (2019).
[10] Chitnis, R., Tulsiani, S., Gupta, S. and Gupta, A.: Intrinsic Motivation for Encouraging Synergistic Behavior (2020).
[11] Devlin, S., Grzeundefined, M. and Kudenko, D.: Multi-Agent, Reward Shaping for RoboCup KeepAway, *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 3*, AAMAS '11, Richland, SC, International Foundation for Autonomous Agents and Multiagent Systems, p. 1227–1228 (2011).
[12] Watkins, C. J. C. H.: Learning from delayed rewards (1989).
[13] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G. et al.: Human-level control through deep reinforcement learning, *nature*, Vol. 518, No. 7540, pp. 529–533 (2015).
[14] Taichi, I., Yoshimitsu, I., Kodai, S., Masato, N., Yusuke, K., Yuki, S. and Chungche, W.: 現場で使える*!Python* 深層強化学習入門強化学習と深層学習による探索と制御, 翔泳社 (2019).
[15] Sutton, R. S., McAllester, D. A., Singh, S. P. and Mansour, Y.: Policy gradient methods for reinforcement learning with function approximation, *Advances in neural information processing systems*, pp. 1057–1063 (2000).
[16] Williams, R. J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning, *Machine learning*, Vol. 8, No. 3-4, pp. 229–256 (1992).