

深層強化学習を用いた 2 on 2 サッカーゲーム AI の報酬

鈴木透弥^{*†} 井上拓哉^{*†} 國府龍之介^{**†} 橘 俊宏^{**†}

[†] 湘南工科大学 工学部 コンピュータ応用学科

1 序論

本稿では深層強化学習 [1] を用いてマルチエージェント環境においてグループの利益を最大化する戦略を行う推論モデルを作成する手法について検証する。学習環境の構成とエージェントアーキテクチャには、Unity エンジンで動作する機械学習フレームワークである Unity ML-Agents[2] (以下 ML-Agents) を用いる。また Unity は機械学習ライブラリ TensorFlow[3] と連携させることができるため、視覚的に学習状況を管理することができるというメリットがある。

本稿では、強化学習を用いたマルチエージェント環境を持つゲームにおいて、複数人のグループに対応するゲーム AI を ML-Agents 2.0 を用いて学習を行い、パラメータとエージェントの挙動の関係について調査を行う。

2 環境の概略

本稿では学習環境としてオープンソースプロジェクト“ML-Agents Toolkit”で提供されている“SoccerTwos”を利用する。4人のエージェントが2対2のおもちゃのサッカーゲームで競う環境である。環境には、2つの異なるマルチエージェントグループがあり、それぞれに2つのエージェントが配置されている。エージェントはチームメイトを観察でき、位置情報から攻守のポジショニングを行う。“SoccerTwos”はエージェント同士が敵対関係にあり、現在と過去の自分を対戦相手として学習する self-play を使用する。エージェントは、ボールが相手ゴールに入ったときにプラスの報酬、自陣のゴールにボールが入ったときにマイナスの報酬、累積時間に応じてマイナスの報酬を得る。

A Maximizing Reward for 2 on 2 Soccer Game AI Using Deep Reinforcement Learning

Touya Suzuki^{*†}, Takuya Inoue^{*†}, Ryunosuke Kokufu^{*†} and Toshihiro Tachibana^{**†}

[†]Department of Applied Computer Sciences, Faculty of Engineering, Shonan Institute of Technology
1-1-25 Tsujido-Nishikaigan, Fujisawa, Kanagawa, 251-8511, Japan

^{*} {19a6061, 19a6011, 19a6049}@sit.shonan-it.ac.jp,

^{**} tachibana@sc.shonan-it.ac.jp

3 シミュレーション実験

3.1 実験 1: 学習回数 max_step の変更

実験 1 では、パラメータを表 1 のようにして max_step = 2.0×10^6 までを行った場合と、max_step 数を半分の max_step = 1.0×10^6 までにした場合について比較する。それぞれ 30 試行を行った際の平均累積報酬を図 1 に示す。図 1 より max_step = 1.0×10^6 の場

表 1: サンプルの提示する基本環境設定

パラメータ名	パラメータ値
trainer_type	MA-POCA[4]
beta	0.005
buffer_size	2048
batch_size	20480
epsilon	0.2
gamma	0.99
hidden_units	512
max_step	2.0×10^6
K_Power	2000
エージェント数	28
ゴールしたときの報酬	1
ゴールされたときの報酬	-1
累積時間報酬	$-1.0/2.0 \times 10^6$

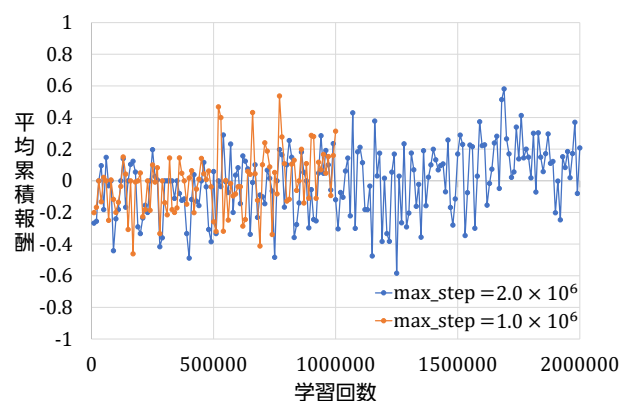


図 1: max_step の違いによる平均累積報酬の比較

合は max_step = 2.0×10^6 の場合よりも、学習に要した時間は短く、より多くの報酬を得ることができた。

max_step = 1.0×10^6 と max_step = 2.0×10^6 の推

論モデルを敵対するよう、グループごとに学習結果を適用して、試合を10回行わせた。試合はどちらかのグループが1点先取した場合を1回とし、1回ごとにボールとエージェントの位置を初期位置にリセットする。

実験の結果、 $\text{max_step} = 1.0 \times 10^6$ が6回、 $\text{max_step} = 2.0 \times 10^6$ が4回先にゴールした。これより $\text{max_step} = 1.0 \times 10^6$ の方が良いエージェントを生成することが出来た。このようになった理由としては $\text{max_step} = 2.0 \times 10^6$ の方が学習時間が長く、 $\text{max_step} = 1.0 \times 10^6$ より累積時間報酬が大きくなり、平均累積報酬が下がった。また、学習回数が増えたことにより、過学習に近い状態になったと推測できる。

3.2 実験2：蹴る力 K_Power の変更

実験2では、エージェントがボールを蹴る力 K_Power を変更した場合について比較する。表1に示す $K_Power = 2000$ と、2倍にあたる $K_Power = 4000$ とした場合の平均累積報酬を比較する。その学習状況のグラフを図2に示す。図2より蹴る力 K_Power を2

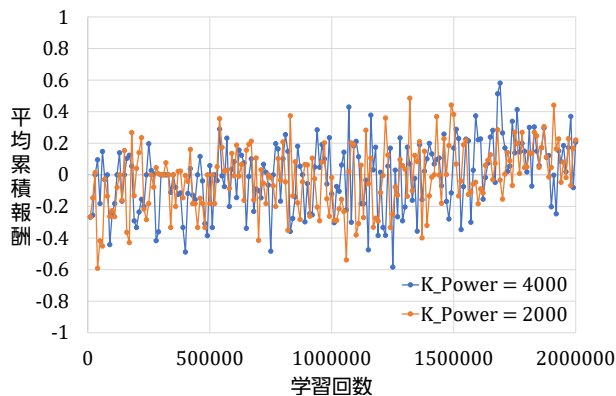


図2: K_Power の違いによる平均累積報酬の比較

倍にすると、 $K_Power = 2000$ の時に比べて学習を終える時間が長くなった。平均累積報酬は $K_Power = 2000$ の時と近い値となった。

また、 $K_Power = 2000$ と $K_Power = 4000$ の推論モデルを敵対するよう、グループごとに学習結果を適用して、試合を100回行わせた。試合は、どちらかのグループが1点先取した場合を1回とし、1回ごとにボールとエージェントの位置を初期位置にリセットする。またボールを蹴る力は $K_Power = 2000$ を適用したグループと、 $K_Power = 4000$ を適用したグループごとに、対応する K_Power の値が適用されるようにエージェントのスクリプトを修正した。

実験の結果、 $K_Power = 2000$ の方が早くゴールした。これは $K_Power = 4000$ と設定した場合、蹴る力

が強くなり、ボールが壁やエージェント自身で跳ね返る回数が増える。この現象により、ボールが予想外の方向へ進んでしまいエージェントが蹴ったボールが壁や敵のグループにぶつかり、オウンゴールを決めてしまう様子が見られた。また、エージェントはボールが減速して停止位置が決まるまでは動きを止める行動が見られ、 $K_Power = 2000$ の条件のときと比較して、ボールを追いかけるまでに時間がかかることがあった。

反対に、表1に示す $K_Power = 2000$ の条件の時は蹴る力が弱いため、エージェントがボールをコントロールしやすく、オウンゴールなどのリスクを減らすことができた。しかし、ボールを遠くまで運ぶことができないため、ボールが途中で止まったところを相手に取られてしまうという場面があった。

4 結論

本稿では、オープンソースプロジェクトの ML-Agents Toolkit で提供されている “SoccerTwos” 環境による深層強化学習を用いて、グループの利益を最大化する戦略を行う推論モデルを作成した。その結果、パラメータの変更によって平均累積報酬と学習時間を変化させることができた。また、self-play は本実験環境で対象としたような、異なるマルチエージェントグループが競争を行う環境下の学習に適していると言える。一方で、本実験環境はエージェントが一家所に固まるような動きが見られ、後方への反応が遅い場面もあった。このような行動を改善するために推論モデルを細かく調査することができない部分もあった。

今後の課題としては、学習の評価軸の検討、ボールの軌道予測、より複雑なエージェント同士の連携、模倣学習、独自の機械学習環境による学習アルゴリズムの検証などが挙げられる。

参考文献

- [1] 布留川英一, “Unity ML-Agents 実践ゲームプログラミング”, ボーンデジタル, 2020.
- [2] Unity Technologies, “Unity Machine Learning Agents Toolkit”, (<https://github.com/Unity-Technologies/ml-agents/>), 2021.
- [3] M. Abadi, *et al.*, “TensorFlow: A system for large-scale machine learning”, *Proc. of OSDI '16*, 265–283, 2016.
- [4] J. Shao, *et al.*, “Credit Assignment with Meta-Policy Gradient for Multi-Agent Reinforcement Learning”, *CoRR*, abs/2102.12957, 2021.