強化学習によるエージェントの戦術獲得の分析

長友結希[1]　三宅陽一郎[1]

概要：本研究では，ゲーム内のエージェント(キャラクター)に強化学習によって自律的に戦術を獲得させる手法の確立を探求する．このような自律的な戦術の獲得は，複雑化するゲーム開発において自動的なキャラクターAIの生成につながるものとして期待することができる．本研究でデモとして提示するのは，対戦型サッカーゲームであり，動的な環境の中でキャラクターが自ら最適な行動のみならず戦術を獲得する結果を提示する．強化学習のフレームワークとしてはUnity ML-Agents に実装された MA-POCA を用いる．強化学習は一般的にゲームスコアを伸ばす方向に開発されるが，本実装ではスコア自体よりもエージェントが新たな戦術を発見することに着目している．また，エージェントの移動ログを解析することで学習を通じてエージェントが戦術を獲得していくことを可視化した．

キーワード：強化学習，マルチエージェント，自己対戦，戦術

Analysis of Acquisition of Strategies for Agents in Reinforcement Learning

YUKI NAGATOMO[1]　YOUICHIRO MIYAKE[1]

Abstract: In this research, we inquire into the establishment of a method for autonomously acquiring tactics by reinforcement learning for agents (characters) in the game. The autonomous acquisition of strategies can be expected to lead the automatic generation of character AI in complicated game development. In this study, we used a competitive soccer game in which the characters acquire not only the optimal behavior but also the strategies. As the framework of the reinforcement learning, we used MA-POCA implemented in Unity ML-Agents. Reinforcement learning is generally developed toward increasing the game score, but this study focuses on the agent discovering new strategies rather than the score itself. In addition, by analyzing the movement log of the agent, it was visualized that the agents acquired tactics through learning.

Keywords: Reinforcement Learning, Multi-Agents, Self-play, Strategies

1. はじめに

本研究では強化学習におけるエージェントの戦術の獲得について研究する．この研究においては，タスクを課された強化学習エージェントがそれを解決する過程で繰り返し似たような方策を取ることを戦術の獲得と見なし，このような戦術を研究の対象とする．戦術の例として将棋では矢倉囲い，バスケットボールではゾーンディフェンスなど，様々な種類のゲームにおいてそのゲーム特有の戦術が用いられている．一般的にはこのような戦術はゲームのプレイを通じて人間が歴史的に発見してきたものであるが，強化学習エージェントも人間のとる戦術と同じような戦術を編み出し，使用することが過去の研究より明らかになっている．AlphaGo Zero [1]においてエージェントは教師なし学習により囲碁を学習したが，その学習過程において複数の定石を自ら見出し繰り返し使用している．その定石の中には既知のものだけではなく，未知のものまであったとされている．また，Capture the Flag [2]では鬼ごっこ状の旗取りゲームをプレイするエージェントがチームメイトを追いかける，相手チームの陣地に張り込む，などの戦術を獲得している．このようなエージェントの戦術獲得について研究することはAIの解釈可能性を向上させうるものであり，作成されたエージェントがどのような状況でどのような方策を取る傾向があるのか評価することができる．エージェントのとる戦術はいわば思考の”癖”のようなものであり，人間に置き換えてみれば「道に迷った時に突き当たりで右折と左折のどちらを選ぶか」のような意識的とも無意識的ともとれるような思考であり，コネクショニズム的なAIと記号主義的なAIをつなげる知見を得るきっかけともなると期待される．本研究はこのような目的のもとでエージェントの戦術獲得について研究を行う．

1 立教大学大学院人工知能科学研究科

Rikkyo University, Graduate School of Artificial Intelligence and Sciences

2. 背景

本研究では環境をシミュレートするにあたり Unity を使用する．Unity は Unity Software Inc. が開発したゲームエンジンであり，3D表現を用いた画像の描画や物理運動の演算を行うことが可能である．Unity を用いることでオブジェクトに重力や摩擦を及ぼしたり，オブジェクト同士の接触の判定をさせたりすることが容易にできるため，物理シミュレーション環境を簡単にシミュレートすることができる．また，Unity 上で使用することのできる強化学習ツールに Unity Machine Learning Agents Toolkit (ML-Agents Toolkit)[3] があり，これは Unity で作成した環境上でエージェントの強化学習を行わせることのできるツールである．ML-Agents にはあらかじめ複数の強化学習手法が組み込まれており，今回の実験では MA-POCA(Multi-Agent POsthumous Credit Assignment)[4] を使用した．MA-POCA は COMA(Couterfactual Multi-Agent) [5] を元に実装されている．以下、COMAとMA-POCAについて簡単に解説する．

(1) Counterfactual Multi-Agent Policy Gradients

COMA はマルチエージェントにおける強化学習を可能にさせる手法である．COMA は Actor-Critic の流れを汲むモデルであり，これをマルチエージェントに拡張した図1のような構造を持つ．各エージェントは個々の観察–行動履歴に基づいて自らの行動を選択するための分散型アクターを持ち，各エージェントから集まる連結行動と環境からの状態を受け取る中央集権型のクリティックを持つ．

ここで Critic はアクター1から行動 u_t^1 およびポリシー $\pi[h^1, \epsilon]$，環境から状態 s_t および報酬 r_t を得て，個々のアクターの反証的価値 $A^1(s, u^1)$ をグローバル報酬 $Q(s, \vec{u})$ とアクター1以外のアクターによる価値との差分として求め，アクターに返す．これにより，各アクターは自分意外のアクターによる状態と報酬への影響を差し引いた学習ができるとされる．

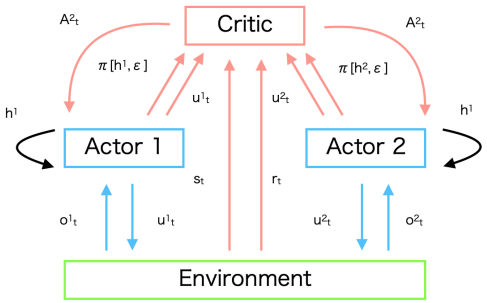

図 1　反証マルチエージェント

Figure 1　Counterfactual Multi-Agent

(2) Multi-Agent Posthumous Credit Assignment

MA-POCAは図2のようにして表されている．Character が個別のエージェントを表し，π がアクター，$o_{1...n}$ がそれぞれの観察，$a_{1...n}$ がそれぞれの行動，Q がクリティックを表している．MA-POCA ではエピソード中のエージェントの増加・減少に対応するためアテンションネットワークが用いられている．

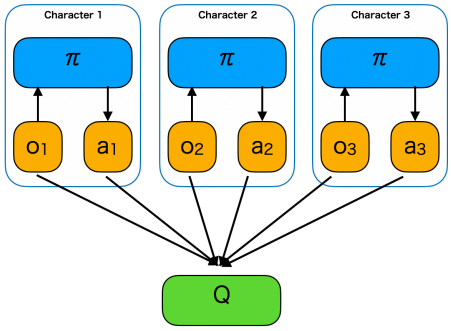

図 2　Multi-Agent Posthumous Credit Assignment

Figure 2　Multi-Agent Posthumous Credit Assignment

3. 実験

3.1. 実験概要

今回の実験では ML-Agents に例として実装されている SoccerTwos を使用した．SoccerTwos は青チームと紫チームの自己対戦型強化学習環境である．各チームには2体ずつのサッカープレイヤーのエージェントがおり，いずれも同じニューラルネットワークモデルを持つ．各エージェントは図3のようにレイキャストと呼ばれるセンサーで環境を観察している．レイキャストはレイをエージェントから放射状に飛ばし，このレイが衝突したオブジェクトに設定されたタグを取得することで視覚的なセンサーとして作用する．SoccerTwos の環境においてレイキャストが取得するタグは{ボール，自チームゴール，敵チームゴール，壁，自チームエージェント，敵チームエージェント}であり，エージェントはこれらを識別できることとなる．エージェントはこのレイキャストから観察として入力を得て，{前進後退，左右平行移動，左右旋回}の行動を出力する．この入力・出力一回分のセットをステップと呼び，このステップを繰り返すことでエージェントはゲームをプレイする．ゲームの一単位はエピソードと呼ばれ，いずれかのチームがゴールをするか，3000ステップが経過することによってエピソードは終了する．ボールはフィールド平面上でしか移動せずフィールドも壁に囲まれているため，ボールがラインアウトすることは無くエピソード終了以外でゲームが止まることはない．エピソードが終了するとボールは環境の中心に再配置され，エージェントもセンターラインから自ゴール側のややランダムに振られた座標に再配置され，次のエピソードが開始される．エージェントは自チームがゴール

することによって正の報酬を得て，敵チームがゴールすることによって負の報酬を得る．また，時間経過に伴って僅かな負の報酬を得る．

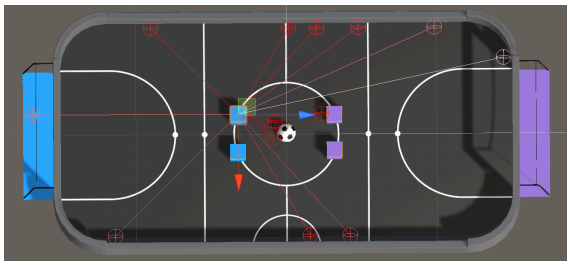

図 3　学習環境におけるレイキャスト

Figure 3　Ray-cast in Learning Environment

3.2. 実験結果

図4は50,000,000ステップの学習を行なったエージェントのイロレーティングを示している．レーティングは1200から開始し，一定ステップ経過ごとにエージェントに割り振るモデルを入れ替え，対戦を繰り返すことで評価を行なっている．イロレーティングはエージェントの性能を表す一定の指標となるが，必ずしもエージェントの行動の変容や戦術の獲得を反映しない．エージェントの行動の変容を評価するために，毎エピソードにおけるエージェントの移動の軌跡を図5のように画像に置き換えて比較を行なった．この比較では10M・20M・30M・40M・50Mステップ学習後のモデルを,学習を行わず推論のみを行う固定されたニューラルネットワークモデル（ONNX ファイル）として出力．これら5つのモデルについてそれぞれ環境上の4つのエージェントに同種のモデルを搭載し，500エピソード分シミュレーションを行いエージェントの座標ログを記録する．それらを t-SNE で2次元へと次元削減を行うことで比較した．図6はこの次元削減を行なった軌跡のグラフと，20Mモデルの各クラスターの軌跡を表すものである．また，図7は同様のグラフと50Mモデルの軌跡を表している．

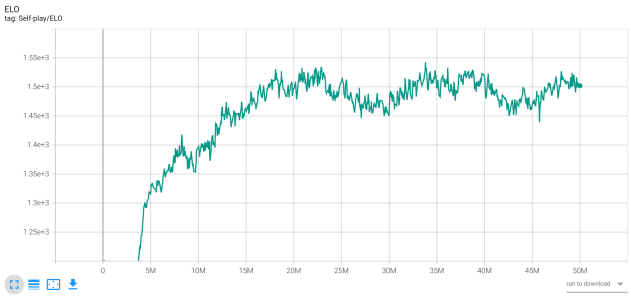

図 4　イロレーティング

Figure 4　ELO Rating

3.3. データ解析と考察

次元削減は概ねモデルごとの違いをクラスター化しており，ほとんどのクラスターは単一のモデルのみから成っている．グラフ中央付近には全てのモデルが密集したクラスターがあり，その軌跡は図6・図7にある通りフィールドのセンターライン付近にわずかにしか残されていない．これは，エピソード開始直後にボールに接触してそのままゴールしているような試合展開となっているためと考えられる．20Mモデルと50Mモデルの違いとしては，20Mモデルは概ねフィールド上の前後の違いが見出しづらい軌跡になっているのに対し，図7左上・左下の軌跡のようにバックコートに軌跡が偏って残されているクラスターが見受けられる．このような軌跡は図8に示すように，青チームの一方のエージェントがフロントコートでオフェンスを行い，もう一方のエージェントがバックコートに残って待機するような行動によって生成される．このような行動は50Mモデルでは比較的よく見られるのに対し，学習時間の短い20Mモデルではあまり見られない．同じ学習時間のものであればフィールド上のどのエージェントも同じモデルで動作しているため，これらのエージェントは個別にオフェンス・ディフェンスを行うように学習したわけではなく，レイキャストによる味方エージェント・敵エージェントとボールの位置関係の観察から判断してオフェンス・ディフェンスを切り替えていると考えられる．そのため，ボールがバックコートに押し込まれオフェンスしていたエージェントとディフェンスしていたエージェントが交差した時などに役割を交代することがある．これは人間のサッカープレイヤーが用いるフォワードやディフェンダーなどのポジション分担による戦術をエージェントが自然に学習したということができるだろう．

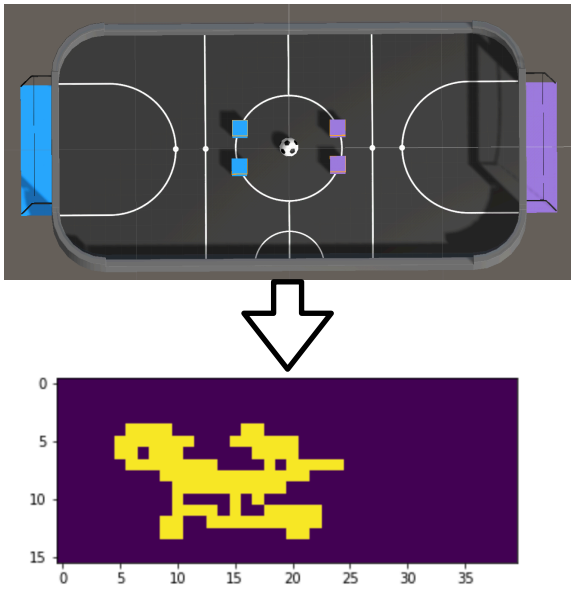

図 5　学習環境と軌跡

Figure 5　Learning Environment and Locus

エージェントの軌跡の t-SNE による次元削減はこのようにモデルごとの行動の違いをクラスター化により明確にすることを可能とし，さらに単一のモデルが使い分ける異なった戦術もまた次元削減によって可視化することができる．今回はエージェントの座標を取得し軌跡を描くことでエージェントの行動を分析したが，その他様々な特徴量からエージェントの戦術が見出せる可能性がある．

4. 結論

本稿では自己対戦型マルチエージェント環境におけるエージェントの戦術獲得について検証を行ない，サッカー型の環境においてエージェントの軌跡を解析することで学習に伴うエージェントの行動変容と戦術獲得を確認できることを明らかにした．この手法はサッカー型の環境のみならず，エージェントの移動が伴う様々な環境に対して適用し得る．

今後の課題としては，異なる環境においてもこの手法を実験することである．エージェントの取りうる行動は必ずしも移動を伴うものではないため，そういった環境では移動とは異なるログを取る必要があるだろう．また，今回の手法でもエージェントが1エピソード内において取りうる戦術が1つのみとは限らず，エピソード単位でログを取る今回の手法では解析できず，何らかの手法でログから該当箇所を切り出さねばならないと予想される．また，今回の手法ではエピソード内の時系列データから時間の要素を抜き取ることによって軌跡に落とし込んでいる．そうすることによってエピソード間の軌跡を比較することができるようにしているが，これによって抜け落ちてしまっている情報も存在する．今後は欠落情報を改善する方法を解決していきたい．

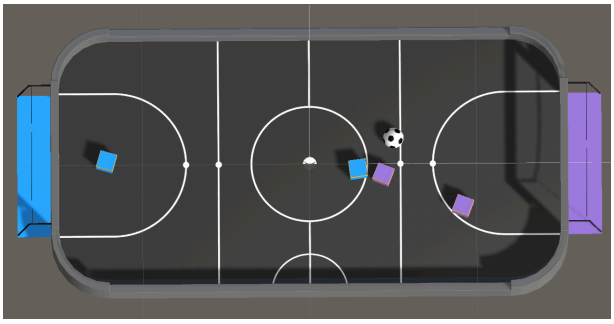

図 8　フォワードとディフェンダー

Figure 8　Forward and Defender

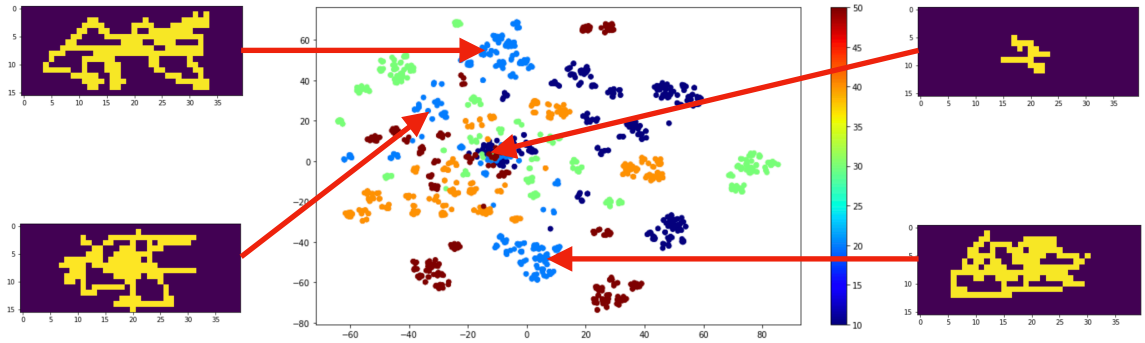

図 6　20M エージェントの軌跡と t-SNE 埋め込み図

Figure 6　20M Agent's Locus and t-SNE Embedding

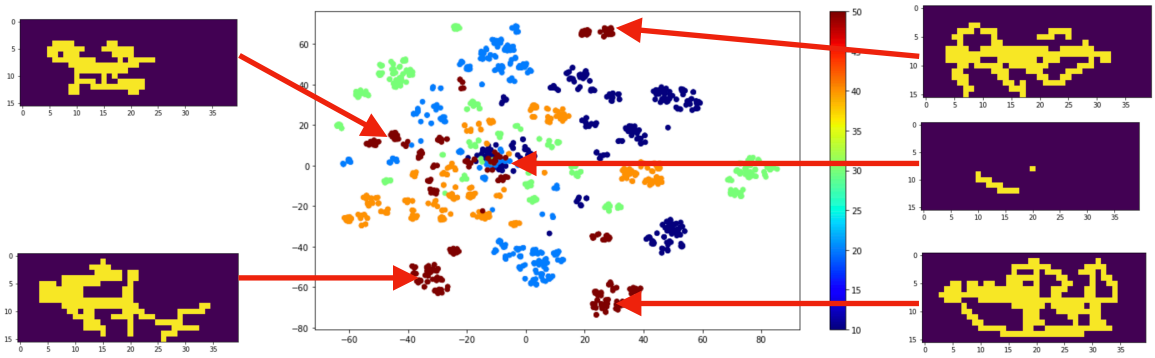

図 7　50M エージェントの軌跡と t-SNE 埋め込み図

Figure 7　50M Agent's Locus and t-SNE Embedding

参考文献

1. Silver, D. et al. Mastering the game of Go without human knowledge. Nature 518, 354-359 (2017).
2. Jaderberg, M. et al. Human-level performance in 3D multiplayer games with population- based reinforcement learning. Science 364, 859-865 (2019).
3. Juliani, A. et al. Unity: A General Platform for Intelligent Agents. https://arxiv.org/pdf/1809.02627.pdf, (参照 2021-10-10).
4. "ML-Agents plays DodgeBall". https://blog.unity.com/technology/ml-agents-plays-dodgeball, (参照 2021-10-10).
5. Foerster, J. et al. Counterfactual Multi-Agent Policy Gradients . Proceedings of the AAAI Conference on Artificial Intelligence, 32(1) (2018).

