# COMA アルゴリズムによるサッカーゲームの戦略的行動学習

永井 宏樹 *

法政大学情報科学部コンピュータ科学科
hiroki.nagai.3g@stu.hosei.ac.jp

## Abstract

Reinforcement learning is a machine learning technique in artificial intelligence in which a learner, called an agent, learns optimal behaviour by trial and error to maximise rewards in an environment where the learner's behaviour can be evaluated. In this study, reinforcement learning is applied in the context of a multi-agent system in the Google Research Football football simulation environment to improve agent performance. Specifically, in this environment where interaction plays an important role, the Counterfactual Multi-Agent Policy Gradients (COMA) algorithm is used to enable individual agents to use centralised Critic and learn optimal action learning to select the best action given the behaviour of other agents. The experiment was conducted by training in different FootBall Academy scenario environments and evaluating the results in terms of scores. The results showed that the maximum reward of a score of 2.0 was obtained in the academy-empty-goal-close scenario, and higher scores were obtained as the number of participants increased. As a factor, we estimate that the addition of the checkpoint reward improved accuracy and created a more accurate agent.

## 1 序論

強化学習とはエージェントと言われる学習者の行動を評価できる環境で、報酬が最大となるように試行錯誤しながら行動し、最適な行動を学習する人工知能の機械学習手法の 1 つである。近年この強化学習を使ったマルチエージェントシステムについて研究が多くされるようになった。また、マルチエージェントシステムとは複数の独立したエージェントが相互に協調し、目標を達成するために協力や競争をするシステムであり、エージェント同士の相互作用が重要である研究分野だ。最近では強化学習を使った囲碁や将棋などの強さや勝率を求めたゲーム AI が注目を集めた。

強化学習は単一タスクの性能や将棋囲碁などの探索問題を解決し成功を納めたが、課題として協力するプレイヤーが多いほど共同行動空間が大きくなる問題や競争する相手が定まっていないため、様々な相手に対して適応しなければならない問題、確率性があることでトレーニングの難易度が向上してしまう問題など解決されていない問題がある。また、マルチエージェントシステムでも、複数エージェントが協調してタスクを達成する必要がある場合、エージェント間の協調が難しいという問題がある。この問題を解決するために、本研究ではマルチエージェントシステムのゲーム環境において協調した戦略的学習を行うことで性能の向上を目標とし、マルチエージェントシステムのゲーム環境は 3 D サッカーゲームに注目する。サッカーゲームとは、スポーツのサッカーと同じルールで学習済みエージェントが 11 対 11 で対戦し、ドリブルやパス、スライディングを用いてゴールを目標にするゲームである。このゲームにはパスやコーナーキック、ゴール、ファウル、オフサイドなどのサッカーゲームのルールに関する様々な学習すべき点やカウンターや細かいパス回しなどの戦略を学習することなど、チームのエージェント同士で協調しながらゴールを決めるための動きを学習するため Unity ML-Agents の教材など様々な教材でサッカーゲームが多く取り上げられる。この研究では、エージェントの性能をあげるために、COMA アルゴリズムを適用し性能向上を目指す。

## 2 関連研究

### 2.1 TamakEri

Google Research Football ではこれまでに様々な研究が行われているが、その中で Kaggle で行われた Google Research Football 環境で実施される大会により入賞した TamakEri というエージェントがある。IMPALA アルゴリズムに似た Actor-Learner アーキテクチャを使用してエージェントをトレーニングし、ニューラルネットワークモデルは様々なエージェントとのシングルプレイゲームのセルフプレイを行うことによってトレーニングした。主なアルゴリズムとしては、オフポリシー分散強化学習を使用して多数のゲームの結果からエージェントのポリシーを効率的に更新する。また、方策勾配アルゴリズムにより、モデルはゲームの結果を予測するための価値と割引スコア報酬を見積もるための報酬、またその 2 つを最大化し、次のゴールを相手よりも先に決めるための期待値を最大化するためのポリシーの 3 つを学習する。また、エージェント

* Supervisor: Prof. Katunobu Itou

はセルフプレイの強化学習によって学習されるが、対戦相手はアクションの繰り返しを無くすことで、決定論的なポリシーを使用することで、訓練されたエージェントは常に自分よりも強い相手と対戦することができ、その相手との勝率を上げるために学習を継続的に行える。

TamakEri は過去に行った行動を記録できる GRU を導入した。これにより勾配消失問題が解決された。また、GRU を導入することによって過去の行動を考慮しながら学習でき、より最適な行動を選択するエージェントを作成できた。トレーニングされたエージェントは Google Research Football が提供している、FootBall Benchmarks の最高難易度であるハード (フル 3000 ステップゲーム) のエージェントに対して、1000〜2000 回の評価で 90 %の勝率に達することができた。

しかし、この研究ではサッカーゲームエージェント 11 人に毎回同じ報酬を与えていることが挙げられる。マルチエージェントシステムの報酬分配方法として、各エージェントが自身のみの目標を最適化しようとする個別報酬や報酬をチーム全体に均等配分するチーム報酬という報酬分配方法がある。しかし、特にチーム報酬のような分配方法では、何もしていないエージェントに対しても報酬を与えてしまうため、学習の妨げとなってしまうことがある。 [1]。

### 2.2 Google Research Football: A Novel Reinforcement Learning Environment

Google Research Football を紹介する論文では、フットボールの試合は短期的なコントロール、ボールのパスなどの概念の学習、および高度な戦略の間で自然なバランスが必要とされるため、強化学習にとって特に挑戦的であるとし、そのための環境を安全で再現性のある方法を提供していることが書かれていた。また、Google Research Football には、高度に最適化されたゲームエンジンや FootBall Benchmarks と呼ばれる、難易度が異なる敵エージェント、FootBall Academy という強化学習シナリオのセットなどが提供されており、Github にオープンソースコードがリリースされている。エンジンである FootBall Engine にはプレイヤーの位置やボールの位置、速度などの意味情報を含む様々な状態の特徴表現からの学習と、未加工の画素状況からの学習が可能になっている。また、ランダム性の影響を調べるために、環境と対戦相手の AI アクションの両方にランダム性がある確率モードとランダム性がない決定論的モードの両方を実行することができ、実行するときにはキーボードやゲームパッドなど様々なプラットフォームに対応しており、ゲームに対する感覚を得られる。また、広く使用された OpenAI Gym API と互換性がある。

FootBall Benchmarks では私たちは Football Engine を使った強化学習研究のためのベンチマークのセットを提供し、手作業で設計された対戦相手に対して違和感のないサッカーゲームをプレイさせられる。Benchmarks には Football Easy Benchmark、Football Medium Benchmark、Football Hard Benchmark の 3 つの種類があり、対戦相手は強さだけが異なる。この強さとは、AI の判断速度などが関係する。参考として、2 つの最先端の強化学習アルゴリズムである DQN と IMPALA のアルゴリズムと各 Benchmarks と対戦した際の平均ゴール差は、Easy の対戦相手は、2000 万ステップの訓練を受けた DQN エージェントによって倒す事が可能だが、Medium と Hard の対戦相手を倒すためには、2 億ステップの訓練を受けた IMPALA などの分散アルゴリズムを必要であることが分かった。Football Hard Benchmark は、大規模に分散された RL アルゴリズムでさえも困難であることが証明された。

FootBall Academy は、エージェントを段階的に学ばせるカリキュラム型学習シナリオで、研究アイデアを調査するための基盤を提供している。Football Benchmarks のような考慮しなければいけないルールが多くある環境に最初から対応できるエージェントを訓練することは簡単ではないため、様々なレベルの難易度を持つ多様なシナリオセットから学べる。これにより、研究者は研究を模索、挑戦することが可能となり、パスやシュートなどの高レベルの概念をテストできる。シナリオの例には、エージェントが無人のゴールから得点する方法を学ぶシナリオやプレーヤー間で素早くパスする方法を学ぶシナリオ、およびカウンター攻撃を実行するシナリオなどがある [2]。

### 2.3 DSDF: Coordinated look-ahead strategy in multi-agent reinforcement learning with noisy agents

2024 年 1 月の最新の研究の DSDF では、マルチエージェント強化学習は交通管理や自動運転、自律制御など多くの協調行動を必要とする分野に適用されており、そのアルゴリズムには分散型のポリシーを学習するものがある。また、分散型ポリシーの学習を向上させるために集中型トレーニング法が活用されている。また、集中型トレーニングおよび分散実行方法 (CTDE:centralized training and decentralized execution ) は協調的なマルチエージェントポリシーで使われることが多いが、このアプローチ方法では全て、エージェントがポリシーの指示どおりに動作することを前提としており、場合によってはエージェントが一貫性のない動作をする可能性がある。つまり、ポリシーによって提案されたアクションとは異なるアクションを実行する可能性がある。例えば、エージェントが負傷してしまい動きや精度に影響を及ぼす場合を考える。そのエージェントを劣化したエージェントと呼ぶが、その劣化したエージェントは短い距離のパスなど狭い範囲の短期的な目標で動作させ、ドリブルやランニングなどの長期的で複雑な目標は健康なエージェントに任せることになる。そのため、全体的な共同ポリシーの報酬を最大化するために、全てのエージェントに対して先読み戦略をリファクタリングし、ポリシーの再調整をする必要がある。また、エージェントごとにポリシーで指定されたアクションとは異なるアクションが実行される確率も異なるため、提案された「深層確率的割引率 (DSDF:Deep Stochastic Discount Factor)」法は、ハイパーネットワークを用いて各

エージェントの割引係数を予測する。これにより、劣化したシナリオでの効果的な協力を促進する。強化学習では各エージェントの割引係数を調整することで戦略のリファクタリングを行うことができるため、劣化したエージェントは低い割引率を使用し、短期的な報酬を計画。一方で、健康で協調的なエージェントはより高い割引率を使用し、長期的な報酬を計画する。この研究では、SMAC、Google Research Football、lbForaging、WaterWorld の 4 つのベンチマーク環境でテストされ、既存の方法と比較して平均報酬が高いことを示した。特に今回扱う Google Research Football では、3_vs_1_with_keeper、Run-to_score_with_keeper、Run_pass_and_shoot_with_keeper の 3 つのシナリオで実験を行い、他手法と比べて平均報酬と勝率で高い成果を示した [3]。

## 3 手法

### 3.1 Counterfactual Multi-Agent Policy Gradients

本研究では COMA アルゴリズムを適用することでマルチエージェントのサッカーゲームでの性能の向上を目指す。COMA とはマルチエージェント強化学習アルゴリズムの一種であり、今回このアルゴリズムを選んだ理由は、後述する反事実ベースラインを使用することで、他エージェントを考慮して、他の行動を取っていた時の価値関数を計算し比較することで、協調した学習を行ったからだ。また、TamakEri のような分散型 Actor-Ctitic に対して、Critic を集中型にすることでパフォーマンスと学習速度に関して優れていたからである。StarCraft という 2 D ゲームで敵と戦う際の移動、攻撃、停止コマンドの制御を行い、3 人や 5 人の数種類の敵キャラクターがいるシナリオと戦闘を行った。また比較するために COMA アルゴリズムを使ったエージェントと他の Actor-Critic 法を使ったエージェントを比較した。結果、勝率に関してはどのアルゴリズムよりも高い勝率、高パフォーマンス、学習速度を出し、一番高い勝率のシナリオでは平均 87 %であった。これにより、集中型 Critic である COMA は他の分散型である Actor-Ctitic よりも優れていることが実験でわかった [4]。

| map | heur. | IAC-V | IAC-Q | cnt-V | cnt-QV | COMA |
|---|---|---|---|---|---|---|
| 3m | 35 | 47 (3) | 56 (6) | 83 (3) | 83 (5) | 87 (3) 98 |
| 5m | 66 | 63 (2) | 58 (3) | 67 (5) | 71 (9) | 81 (5) 95 |
| 5w | 70 | 18 (5) | 57 (5) | 65 (3) | 76 (1) | 82 (3) 98 |
| 2d3z | 63 | 27 (9) | 19 (21) | 36 (6) | 39 (5) | 47 (5) 65 |

**表 1.** COMA 実験結果

アルゴリズムは、Actor-Ctitic 法を使っており、行動を実行する Actor とその行動を評価する Critic がある。Critic は Actor から行動やポリシー、環境 (GRF) からは報酬や観測情報をもらい、3 層の Linear 層で報酬予測誤差が最小化されるように価値関数を学習する。価値関数を計算後、他のエージェントの行動を考慮した計算をする反事実ベースラインによって各エージェントの全体に対する貢献度を計算する。この貢献度とアクションを記録する GRU をもとに Actor ではポリシーを更新していくことで、他のエージェントの行動を考慮しながら学習する。反事実ベースラインとは、特定のエージェントの全体に対する寄与度を計測すれば特定エージェントが取った行動の全体に及ぼす価値 Rt を計算できるといった考えから、価値関数と特定のエージェントが協調に反する行動を起こした場合の価値関数である反証的価値関数との差を使った差分報酬によって、特定エージェントがとった行動が全体にどれくらいの価値があったかを計算する。

$R_t = A^a(s, \vec{u}) = Q(s, \vec{u})–bs(s_t)$

Rt はステップにおけるエージェント a の報酬を表し、エージェントの行動選択がどれだけタスクに貢献したかを示す。Q(s,u) は状態 s と行動ベクトル u に対する行動価値関数であり、特定の状態で特定の行動ベクトルをとった場合の期待報酬を評価した。bs(st) は特定のエージェントが行動した時の期待値である。

$\mathrm{bs}(\mathrm{s}_t) = \sum_{\acute{u}^a} \pi^a(u^a|\ \tau^a) Q(s, (\vec{u}^{-a}, \acute{u}^a))$

この変数は特定エージェントの他の行動を取った時の価値関数を表しており、式の中で総和を取ることで他のエージェントが可能な全ての行動について考慮でき。また、エージェント a が行動 (u') を選択する確率を表す。エージェントのポリシーに基づき、状態 (s) と過去の経験をもとで行動が選択される。状態 (s) と他のエージェントの行動を含む全体の行動ベクトルに対する行動価値関数であり、そのときエージェント a が選択した行動の期待報酬を評価する。それぞれの行動が報酬にどれだけ貢献したかを求めることができ、他のエージェントの行動に適切に対処することで効果的な協調した戦略的学習ができる。また、この利点と集中型 Critic による性能向上を目指す [5]。

### 3.2 CheckPoints 報酬の追加

報酬は、ゴールにボールを入れることで得ることができる Scoring 報酬と、敵ゴールからのユークリッド距離を参考に、どれだけ敵ゴールに近づくことができたかという Checkpoints 報酬の 2 つを与えることで学習を行い、これを最大化することを目的とする。従来研究では Scoring 報酬のみであるが、Scoring 報酬のみであると学習速度が遅く、ゴールに入れることで報酬を得られることを学習することも多くの学習ステップが必要であるためである。そのため、Checkpoints 報酬を追加することで、Scoring 報酬を得やすくすることを目的に追加を行う。Scoring 報酬はゴールに入れることで + 1 の報酬を受け取り、逆に入れられると- 1 の報酬を受け取る。Checkpoints 報酬は敵ゴールに近づくほど大きい報酬を得ることができ、最大 1.0 の報酬を受け取ることができる。また、Checkpoints 報酬はボールを持ったエージェントが敵のゴールにどこまで近づくことができたかで報酬を与え、最大 1.0 の報酬を受け取れる。

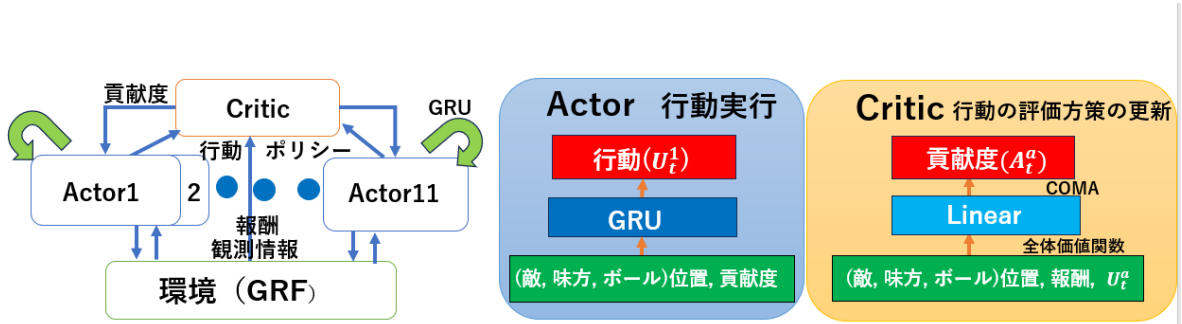


**図 1.** COMA アルゴリズム

## 4 評価

### 4.1 実験方法

作成したモデルを用いて、Google Research Football が提供する FootBall Academy のシナリオにおけるマルチエージェント強化学習でのスコア向上を評価する。

また、100 回の試行の平均スコアで計算を行う

FootBall Academy にある以下の 4 つのシナリオで実験を行いスコアで評価を行う。

- academy_empty_goal_close
  無人のゴールに対して、ゴールにいれるシナリオ
- academy_run_pass_and_shoot_with_keeper
  2 人の選手がパスをしながら相手キーパーがいるゴールへの得点を試みる。1 人はボールを持っておりでマークされておらず相手キーパーと向き合っている。もう 1 人は中央で相手ディフェンダーの隣に配置されている。
- academy_3_vs_1_with_keeper
  3 人の選手がパスをしながら相手キーパーがいるゴールへの得点を試みる。1 人は各サイドに、もう 1 人は中央に配置されている。中央のプレーヤーがボールを持ち、相手ディフェンダーの方を向いている。
- 5_vs_5
  キーパーを含む 5 対 5 のフルゲームサッカーゲームで試合を行う。

**図 2.** academy_run_pass_and_shoot_with_keeper

シナリオでは、相手ゴールにボールを入れたら報酬が +1、逆に味方ゴールに入れられたら-1 となる Scoring 報酬とエピソードに 1 回相手ゴールからのユークリッド距離を用いて、敵のゴールに近づけるほどに報酬を最大 +1 もらえる CheckPoints 報酬の 2 つの報酬を与えている。academy_empty_goal_close は最も簡単なシナリオであり、敵エージェントなど、障害となるものを省いた環境で学習ができるかを評価する。academy_run_pass_and_shoot_with_keeper、academy_3_vs_1_with_keeper の環境では、複数人のエージェントがパスを出しながら敵ゴールにシュートを決めるものである。この実験によって、味方の動きを考慮した行動を取りながらゴールを決めることを目指す。パスを細かく行うことで、相手にボールを取られにくくし、ゴールを決めることについて比較を行い、人数によって報酬が変わるかを評価する。5_vs_5 では、キーパーを含めた 5 人の環境でトレーニングを行い、評価を行う。他 3 つのシナリオではゴールを決める、相手にボールを取られる、コート外に出るなどが起きた場合エピソードが終了するが、5 対 5 のシナリオでは、フルゲーム (3000 ステップ) の 5 分間の試合を行う。

### 4.2 結果

実験の結果、表 2 のような結果が得られた。

表 2 より、academy_empty_goal_close に関しては Scoring + Checkpoints 報酬では受け取ることのでき

| senarios | 従来研究 Scoring | Scoring | Scoring + Checkpoints |
|---|---|---|---|
| empty_goal_close | 0.95 | 1.00 | 2.00 |
| run_pass_and_shoot_with_keeper | 0.84 | 0.12 | 0.70 |
| 3_vs_1_with_keeper | 0.89 | 0.36 | 0.86 |
| 5_vs_5 | -3.00 | -1.02 | -0.26 |

**表 2.** 各シナリオでのスコア

る最大値の 2.0 を獲得することができ、Scoring のみの結果も最大値である 1.0 を獲得することができた。また、academy_run_pass_and_shoot_with_keeper と academy_3_vs_1_with_keeper では、Scoring 報酬を受け取ることができたため、ゴールを決めていることがわかる。また、Scoring + Checkpoints の報酬では約 0.8 のスコアを獲得することができた。5_vs_5 では、スコアが負のスコアになっているため、対戦相手に負け越していることがわかる。

従来研究と比較すると、academy_empty_goal_close では従来研究が 0.95 であることから、精度の高さを示せた。しかし、run_pass_and_shoot_with_keeper と 3_vs_1_with_keeper では、従来研究よりも低い精度であることがわかる。5_vs_5 では、負け越しているが、従来研究 [6] よりは失点率が低く相手にゴールを決められない方向に精度を高めることができた。

また、マークされていない味方に対してパスを出すことが多く学習することができた。

以下に各シナリオの学習曲線を示す

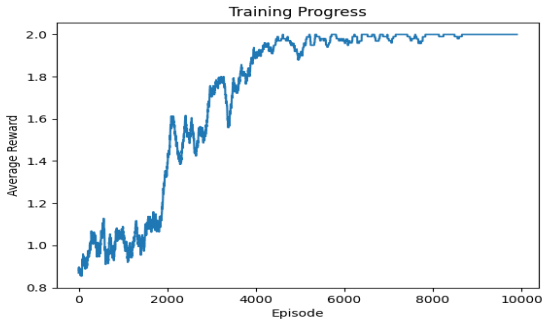


**図 3.** academy_empty_goal_close の学習曲線

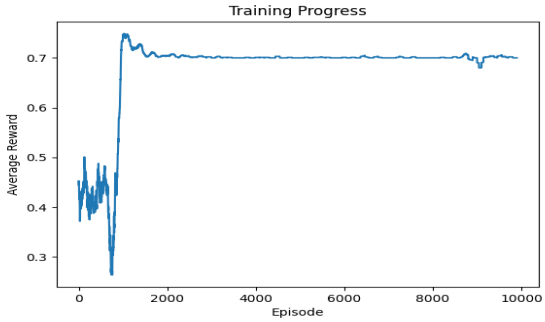


**図 4.** academy_run_pass_and_shoot_with_keeper の学習曲線


**図 5.** academy_3_vs_1_with_keeper の学習曲線

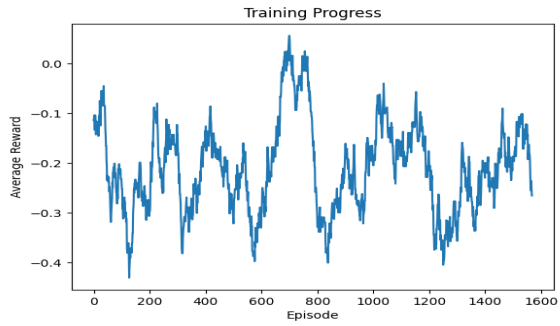


**図 6.** 5_vs_5 の学習曲線

academy_empty_goal_close は最大値である 2.0 まで学習することができ、最後まで安定して学習できた。しかし、academy_run_pass_and_shoot_with_keeper と academy_3_vs_1_with_keeper は約 0.8 で学習が止まってしまっていることがわかる。5_vs_5 では、3000 ステップゲームを 500 万ステップでトレーニングを行った。その結果、学習が進まなかった。原因として、他のシナリオと比べてステップ数が多かったため、学習回数が少なかったことが考えられる

## 5 考察

本研究では、Google Research Football における FootBall Academy の複数シナリオ環境において、COMA を適応し、Checkpoints 報酬を追加することによってスコアの評価を行った。その結果、academy_empty_goal_close の一番簡単な環境において、そのシナリオで受け取ることができる報酬の最大値である 2.0 を獲得し、Scoring 報酬のみでも 1.0 と最大値を獲得できた。この結果は、従来研究 [2] と比較して高い精度を示せた。これは、Checkpoints 報酬を追加したために

ゴールに近づくことを学習したため、精度が上がったと考えられる。しかし、academy_run_pass_and_shoot_with_keeper と academy_3_vs_1_with_keeper に関しては、Scoring 報酬を受け取ることはできたものの従来研究 [3] と比べて低い値となってしまった。この要因としては学習ステップ数の差と図 7、図 8 より Checkpoints 報酬が挙げられる。academy_run_pass_and_shoot_with_keeper では 0.7、academy_3_vs_1_with_keeper では 0.85 程度で学習が停滞していることがわかる。これは、毎エピソード与えれる Checkpoints 報酬を最大化しようとして、報酬が稀な Scoring 報酬が薄れていることが考えられる。これを解決するためには、2 つの方法が考えられる。1 つは報酬のスケーリングである。FootBall Academy では「ゴールをする」ことが目標であるため、Scoring 報酬が重要である。そのため、Checkpoints 報酬を小さくすることで、Scoring 報酬が重要であることを学習させる。実験として、Checkpoints 報酬が与えられるときに 0.1 倍にした。しかし、期待した結果は出なかったため、さらに小さくすることで Scoring 報酬に焦点をあてることができるのではないかと考える。2 つ目として、独自の報酬を追加することである。例えば、Scoring 報酬を得ることが出来ずに終了してしまったときに、-0.01 のようなマイナスの報酬を追加することや味方にパスが通ったときに 0.01 などのプラスの報酬を追加することが考えられる。また、Checkpoints 報酬を使用したポジショニングを行うことは戦略的な学習において有効であると考える。例えば、ディフェンダー役が守備ラインを維持する位置やミッドフィルダー役が中央を制御する位置、フォワード役が敵ゴールに近い位置に目標位置を設定しポジショニングをする。また、Academy ではセンタリングする位置に目標位置を設定し、的確にパスを出すことでスコアを上げられるのではないかと考えられる。今後の課題として、報酬が少なくとも 1.5 以上が望ましいため、これらの方法を試し Scoring 報酬を高める必要がある。また、5_vs_5 では、負のスコアになっているため相手に負け越していることが考えられ、学習が進まなかった。原因として、他のシナリオと比べてステップ数が多かったため、学習回数が少なかったことが考えられる。しかし、従来研究 [6] と比較すると、失点率は低いことがわかる。このことから、提案手法ではパスなど協調した行動をとることによって失点を防ぐことができたと考えられる。

COMA を使ったトレーニング、実行では academy_run_pass_and_shoot_with_keeper と academy_3_vs_1_with_keeper のように同じ環境で人数が増えるとスコアが上がっていることがわかる。これは COMA が他のエージェントの行動を考慮して協調した行動を学習できたと考えられる。11 対 11 のフルゲームなどで十分なステップ数でのトレーニングに拡張することが課題である。

## 6 結論

本研究では、Google Research Football 環境で COMA アルゴリズムを適用し、報酬に Checkpoints を追加することで性能向上を目指した。結果として、academy_empty_goal_close では高いスコアを得ることができたが、academy_run_pass_and_shoot_with_keeper と academy_3_vs_1_with_keeper では従来研究よりも低いスコアになってしまった。また、5_vs_5 のシナリオでは負け越してしまったものの従来研究 [6] よりも失点率低くすることができた。その理由として、COMA で他のエージェントの行動を考慮する学習ができたことによって、マークされていない味方に対してパスを出すなどの行動を取れたからである。しかし、academy_run_pass_and_shoot_with_keeper と academy_3_vs_1_with_keeper などのように、Checkpoints 報酬を最大化してしまうなど、まだ報酬設計などに課題が残る。

## 参考文献


[1] Katsuki Ohto. Tamakeri. https://www.kaggle.com/c/google-football/discussion/203412, 2020.

[2] Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zajac, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, and Sylvain Gelly. Google research football: A novel reinforcement learning environment, 2020.

[3] Satheesh Kumar Perepu, , Kaushik Dey, and Abir Das. Dsdf: Coordinated look-ahead strategy in multi-agent reinforcement learning with noisy agents. *Proceedings of the 7th Joint International Conference on Data Science Management of Data (11th ACM IKDD CODS and 29th COMAD)*, Vol. 24, pp. 73–81, 2024.

[4] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients, 2017.

[5] Matteo Karl Donati. Counterfactual-multi-agent-policy-gradients. https://github.com/matteokarldonati/Counterfactual-Multi-Agent-Policy-Gradients/blob/master/training*oop.py*, 2020.

[6] Jingqing Ruan, Yali Du, Xuantang Xiong, Dengpeng Xing, Xiyun Li, Linghui Meng, Haifeng Zhang, Jun Wang, and Bo Xu. Gcs: Graph-based coordination strategy for multi-agent reinforcement learning, 2022.