修 士 論 文 の 和 文 要 旨

研究科・専攻	大学院　情報理工　学研究科　　情報学　　専攻　博士前期課程		
氏　　名	阿部 宇志	学籍番号	1930006
論 文 題 目	サッカータスクの深層強化学習における段階的な協調行動の獲得		

要　旨

本研究では，サッカータスクでの協調行動を獲得するため，段階的に学習を行うアプローチであるカリキュラム学習を用いて研究を行った．サッカータスクは報酬がスパースなタスクであるため，どんな行動が報酬や罰につながるかを明確にすることが難しく，状態の多さや行動の複雑さから学習が困難になる．そのため，複雑なタスクであるサッカータスクにおいて，サッカータスクに適したアプローチを用いる必要がある．本研究では，簡単なタスクから学習を始め，徐々に難しいタスクを学習させるカリキュラム学習で協調行動獲得の学習の効率化を目指した．これまでのサッカータスクにおけるカリキュラム学習は協調行動の獲得に関して研究が行われてこなかった．カリキュラム学習で協調行動を学習させるために，本稿では人が行うサッカーの練習に似せて，コーンのような障害物を敵に見立てることや段階的に敵のエージェントを増やすことで，タスクを難しくしてカリキュラム学習を行った．実験ではカリキュラム学習を促すために行った Reward Shaping の効果をみる Reward Shaping における実験と，カリキュ ラム学習の効果をみる実験を行った．前者では，2 体のエージェントの協調行動が必要な環境で実験を行い，Reward Shaping を行った場合が行わなかった場合を目標達成率で上回った．後者では，シュートチャンスでの 2 体のエージェントの協調行動を試みた学習を行い，カリキュラム学習をした場合がカリキュラム学習をしなかった場合に比べて，目標達成率を上回ることを示した．

令和 2 年度修士論文

サッカータスクの深層強化学習における段階的な協調行動の獲得

電気通信大学 大学院情報理工学研究科 情報学専攻
メディア情報学プログラム

学籍番号 : 1930006
氏名 : 阿部 宇志

主任指導教員 : 大須賀 昭彦 教授
指導教員 : 田原 康之 准教授
指導教員 : 清 雄一 准教授

提出年月日 : 令和 3 年 1 月 25 日 (月)

概要

本研究では, サッカータスクでの協調行動を獲得するため, 段階的に学習を行うアプローチであるカリキュラム学習を用いて研究を行った. サッカータスクは報酬がスパースなタスクであるため, どんな行動が報酬や罰につながるかを明確にすることが難しく, 状態の多さや行動の複雑さから学習が困難になる. そのため, 複雑なタスクであるサッカータスクにおいて, サッカータスクに適したアプローチを用いる必要がある. 本研究では, 簡単なタスクから学習を始め, 徐々に難しいタスクを学習させるカリキュラム学習で協調行動獲得の学習の効率化を目指した. これまでのサッカータスクにおけるカリキュラム学習は協調行動の獲得に関して研究が行われてこなかった. カリキュラム学習で協調行動を学習させるために, 本稿では人が行うサッカーの練習に似せて, コーンのような障害物を敵に見立てることや段階的に敵のエージェントを増やすことで, タスクを難しくしてカリキュラム学習を行った. 実験ではカリキュラム学習を促すために行った Reward Shaping の効果をみる Reward Shaping における実験と, カリキュラム学習の効果をみる実験を行った. 前者では, 2 体のエージェントの協調行動が必要な環境で実験を行い, Reward Shaping を行った場合が行わなかった場合を目標達成率で上回った. 後者では, シュートチャンスでの 2 体のエージェントの協調行動を試みた学習を行い, カリキュラム学習をした場合がカリキュラム学習をしなかった場合に比べて, 目標達成率を上回ることを示した.

目次

図目次

第 1 章

はじめに

1.1 背景

1.1.1 マルチエージェント強化学習

深層学習と強化学習を組み合わせた深層強化学習は, ゲーム AI[1][2] やロボット制御 [3][4] などで広く使われてきた. ただし, 現実世界では, 1 体のエージェントが単独で解決できるタスクだけでなく, 複数のエージェントによる協調的な行動が求められることがある. 複数のエージェントが関わる強化学習は, マルチエージェント強化学習と呼ばれ, 研究が行われている.

マルチエージェントタスクは, ライドシェアリングシステム [5] や信号制御 [6] など, 現実世界でも多く存在する. Li らは, ライドシェアリング問題におけるオーダーディスパッチ問題を扱い, シミュレーションにおける累積運転者所得（ADI）と注文応答率の指標で, いくつかの強力なベースラインよりも優れた性能を示した [5]. Prabuchandran らは, マルチエージェント強化学習アルゴリズムを適用して, 動的な交通制御方策を獲得し, それぞれの交通信号を独立したエージェントとしてモデル化させ, VISSIM でのシミュレーションでは, これまでのアルゴリズムよりも優れた性能を示した [6]. これらのタスクからみれるように, 強化学習を現実世界で活用するための様々なマルチエージェント強化学習におけるタスクで研究が行われている.

1.1.2　サッカータスクの強化学習

近年, マルチエージェント強化学習の研究の中で, サッカータスクが多く用いられてきた. これまで, 実際のロボットを動かしてサッカーを行う RoboCup サッカー [7] やシミュレーションを行う Google Research Football[8], 物理演算エンジン MuJoCo[9] のサッカータスク [10] など様々な手段で研究が行われてきている. サッカータスクは, 得点時や失点時にしか報酬や罰を得ることができないため, どんな行動が報酬や罰に関わるかをエージェントが明確に認識することが難しい. こういった報酬がスパースなタスクでは学習が困難になることが知られている [11]. またサッカータスクでは, エージェントが単独で得点したり, 単独で失点を防ぐことが非常に難しいため, 場面に応じて適切な協調行動や敵対行動をとる必要がある. こういった複数のエージェントが関わる必要があるタスクでは, 味方や敵の行動を考慮して最適な行動を選択しなければならない. そのため, エージェントが考慮すべき情報量が多くなり, 探索に時間がかかることから学習が困難になる.

1.1.3　カリキュラム学習

本研究では, サッカータスクでの協調行動獲得のために, 段階的な行動獲得のアプローチとしてカリキュラム学習 [12] を用いて研究を行った. Bengio らによって定義されたカリキュラム学習は, 学習が難しいタスクを解決するために, はじめは簡単なタスクから学習し, 徐々に難しいタスクを学習する手法である. この研究では, 自動生成された図形の分類するタスクや, 単語列から次の単語を推定するタスクで, カリキュラム学習の有効性を示している. 人はなにかを学習するとき, はじめから難解な問題に取り掛かるのではなく, 簡単な問題から徐々に難しい問題を学習する. エージェントの学習においても, 人と同じようにカリキュラムを組み立てることで学習を効率よく進められるようになる.

これまで, サッカータスクでのカリキュラム学習の研究は行われてきたが, サッカーに必要な協調行動の獲得を学習することはできていなかった. 本研究では, サッカータスクの協調行動をカリキュラム学習を用いて獲得するため, 動かない相手をコーン（サッカーの練習で用いられる障害物）として見立てて学習し, 段階的に相手を増やしたり, 守備エージェントに動きを加えることで効率的な学習を促す. 本稿では, シュートチャンスでの協調行動を想定した環境で実験を行い, カリキュラム学習を行った場合のゴール率がカリキュラム学習を行わなかった場合を上回った.

1.2　論文の構成

本論文の構成は以下のとおりである．1 章で研究背景を述べ, 2 章で関連研究, 3 章で本研究での提案手法を説明する. 4 章で実験結果を示し, 5 章で考察, 6 章で本研究でのまとめを述べる.

第 2 章

関連研究

2.1 サッカータスク

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 [13] 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, Google Research により開発された Google Research Football[8] での研究も行われている. Google Research Football は, テレビゲームのようにパスやシュート, 選手の移動などがすべてボタン入力で行われ, ファールやオフサイドなど現実のサッカーと同じ設定で試合が行われているシミュレータである. また, 状態の表現が生データやそのほか様々な表現方法で表されており, それらのデータをもとに学習を行うことができるプラットフォームである. 近年では, Kaggle 社によってコンペティションが開催され, そこではルールベースのアプローチや機械学習アプローチ, 強化学習アプローチなど多くの研究者によって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 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 [10] 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. Chitinis らも, MuJoCo のサッカータスクで 2 体のエージェントがボールに関わってゴールするタスクで実験を行い, 予測した状態との違いを内発的報酬として外部報酬に加え

て学習することで, 報酬がスパースなタスクにおける協調行動獲得を促している. これらのように, サッカータスクにおける強化学習の研究は多く行われている.

2.2 Reward Shaping

報酬がスパースなタスクでは, エージェントが学習中に最終目標に到達するためのヒントを得ることが難しく, 学習の効率が悪くなるため, 様々なアプローチで学習の効率化が研究されている. このような報酬がスパースなタスクに対するアプローチの一つとして, エージェントが段階的に報酬を受け取ることができる Reward Shaping[14] が提案されてきた. Reward Shaping 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. 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. 例えば, ペットボトルの蓋を開けるタスクの場合, 蓋をひねるだけではペットボトル全体が回転してしまい, 蓋を開けることが出来なくなってしまう. そのため, ペットボトルの蓋を開けた時に与えられる外部報酬だけでなく, 逆の手でペットボトルを握ったり, 蓋をひねるという行動をしたときに Shaping 報酬をエージェントに与えることで, 学習を比較的容易に誘導することできる. 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 [15] 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.

しかし, Reward Shaping では Shaping 報酬が不適切に設定されている場合, 学習に悪影響が出てしまう [14]. 例えば, ペットボトルを開けるとき, 蓋をひねるべき方向と逆の方向にひねったときに Shaping 報酬が与えられると, 誤った誘導になり, 学習の妨げとなってしまう. したがって, そのタスクにおけるドメイン知識を持つ設計者によって, 適切な報酬設計が行われる必要がある.

2.3　カリキュラム学習

また, サッカータスクの強化学習を効率的に進めるアプローチの一つに, カリキュラム学習を用いた研究がある. Narvekar ら [16] は, カリキュラムを構築する有用なソースタスクの生成方法に焦点を当てた研究を行い, RoboCup タスクである Half Field Offense タスク [17] とパックマンドメインで提案されたカリキュラム学習を行うことで, 学習を加速させることを示した. また, Silva らの研究 [18] でもサッカータスクのカリキュラム学習として, タスクの生成とカリキュラムの自動生成を行い, Half Field Offense タスクと作成された Gridworld タスクとで学習の効率化を示した.

しかし, これまで行われたサッカータスクでのカリキュラム学習の研究 [16][18] では, サッカーに必要な協調行動を学習することはできていない. Silva らの研究では, 攻撃エージェント 1 体に対して守備エージェント 2 体の環境で攻撃エージェントの学習が行われており, Narvekar らの研究では攻撃エージェント 2 体（ボールを保持しているエージェントのみ学習）に対して守備エージェント 1 体の環境で学習が行われている. 複数エージェントの協調行動の学習は, 味方の行動の意図を考慮しない単独のエージェントの学習に比べ, 味方の動きを予測することも必要になるため, 学習が難しくなる. パス交換や組織的な守備など実際の試合でも多くの場面で必要なスキルである協調行動の獲得は, サッカータスクでの学習において欠かせない目標であるが, これを解決することはこれまでできていない.

また, これまでのサッカータスクでのカリキュラム学習の研究 [16][18] ではドリブルやパス, シュートなどのようにあらかじめ設計されたスキルを用いて行動を学習している. ロボットにおける処理は, 低レベルの行動制御から協調行動や役割分担まで幅広いレベルの処理を含む [19]. サッカータスクのような動作が複雑なタスクでは, 体のどの部分に力を入れ, どの順序で体を動かせば走ることができるのか, またボールを蹴ることができるのかなど低レベルの行動から学習する場合, あらかじめ設計されたスキルを用いて学習する場合に比べて学習が困難になることが知られている. 低レベルからの行動獲得は, 事前に行動を設計する手間を省くことができ, 設計者による開発技術の優劣がエージェントごとにつきにくくなるという利点があるが, これまでのサッカータスクでのカリキュラム学習では, 低レベルから行動を学習している研究は著者の知る限り行われていない.

本研究ではこれらの課題を含んでいるサッカータスクでのカリキュラム学習による行動獲得に対して, 特にカリキュラム学習を用いた協調行動の獲得に重きを置いて, 限定されたシュートチャンスでの攻撃における行動を学習する. サッカータスクでの協調行動の獲得は, 様々な場面で必要とするが複数のエージェントが行動を行うため, 学習が困難になる. しかし, これまでのサッカータスクでのカリキュラム学習では, 単独のエージェントのみの行動獲得だったため, 本研究ではカリキュラム学習としてコーンを用いて 2 体のエージェントを同時に学習させ, 協調行動の獲得を行うこととした. また, 低レベルの行動からでもカリキュラム学習によって協

調行動を学習できることを示せるように, 物理演算エンジン MuJoCo[9] のサッカータスク [10] であらかじめ行動を設計せずに実験を行うことで, カリキュラム学習の有効性を示す.

第 3 章

提案手法

3.1 問題設定

本研究では, 報酬がスパースなサッカータスクでの獲得が困難な協調行動を学習するために, 段階的に学習を行うアプローチとしてカリキュラム学習を用いた研究を行う. 実験では, サッカーシミュレータとして物理演算エンジン MuJoCo[9] のサッカータスク [10] を用いて, 図 3.1 で表されるシュートチャンスでの協調行動を獲得する.

この環境では, 攻撃側のエージェントは守備を学習した守備エージェントをかわしてゴールを決める必要があり, 離れた場所にいる味方のエージェントにパスを出してからゴールすることが効率よく, エージェント間の協調行動が求められる状況になっている. サッカーでは, このようにシュートチャンスでもパスを選択したほうが得点につながりやすい場面がある. そのため, 場面に応じて味方と敵, ボールやゴールの位置など考慮すべき情報を正しく処理して行動しなければならない. 図 3.1 のようなシュートチャンスやその他の連携のとれた守備など協調行動が求められる場面が多いが, エージェントにとって考慮すべき情報が多いため学習は難しくなる.

また, 段階的な学習において低レベルから行動を学習させるように, 実験では 3 次元の値で表現された行動で学習を行う. 3 次元の値はそれぞれ, 体を前後を加速させる値, 体を回転させる値, 下向きの力を加えてジャンプさせる値で示されており, 図 3.1 の環境で協調行動を学習するには, まずボールをどのようにしたら蹴ることができるか学習することから始める必要がある. 本研究では, 守備エージェントに奪われにくいドリブルや指定されたエージェントへのパスなどのスキルはあらかじめ設計せず, その状況に適した 3 次元の値をエージェントが出力できるように学習を行う.

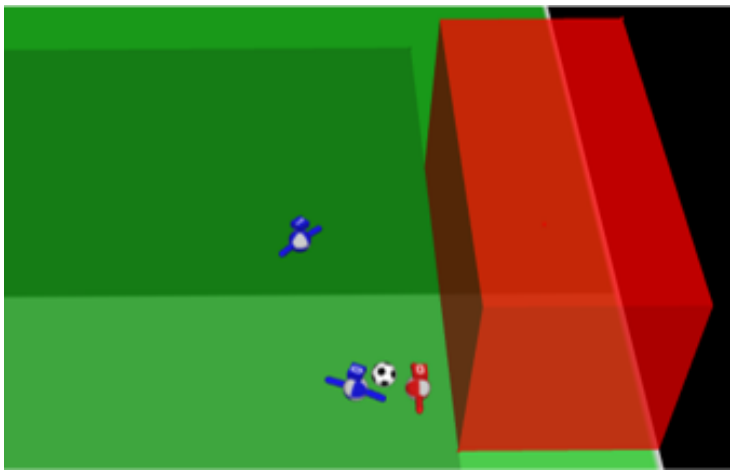


図 3.1　問題設定

3.2　アプローチ

3.2.1　カリキュラム学習

本研究では, コーンを用いたカリキュラム学習 [12] でのエージェントの協調行動獲得を促す. 人がサッカーの練習を行うとき, 例えばコーンを相手に見立ててシュート練習をした後に, 実際に守備選手をつけてシュート練習を行うことがある. ある特定のシーンを想定して, コーンを使って段階的に難しい練習を行うことで, スムーズに感覚をつかむことができる. そういった練習方法に倣い, 本研究ではまず守備エージェントがいない状態でのパス交換を学習し, 段階的に守備エージェントを増やしてカリキュラム学習を行う.

カリキュラム学習の全体像を図 3.2 に示す. まず環境 1 のように敵がいない状態, 敵が 1 体いる状態で環境を用意し, 敵がいない状態で学習させた学習済みモデルを敵が 1 体いる状態に継承させることでカリキュラム学習を行う. ここまでは敵のエージェントを動かさず, コーンのように見立てて攻撃側にとって容易な環境で学習を行う. また, 2 体のエージェントがボールに関与してゴールができれば得点という条件をつける. もし 1 体だけでゴールを決めたとしてもゴールは認められず, 協調行動が不可欠な環境を作っている. カリキュラム学習でエージェント同士がパスをしてからゴールすることを促すことで, 本来の目的である環境 2 での目標達成率を高めることを促す.

環境 2 はこれらで学習された学習済みモデルを用いて, 守備を学習した敵 1 体を相手にして学習を行う. 守備エージェントは, 単純にゴールを決めるだけの行動を学習した攻撃エージェントに対して, ゴールを奪われないように学習させたエージェントを使用している. 環境 2 では, これまでのような得点における条件をつけずに, エージェントがゴールを決めれば得点とする. ゴールに向かってボールを蹴るだけでは, 守備エージェントによってゴールを防がれてしまうため, ボールを奪いに来る守備エージェントをかわしてゴールする必要がある. 環境 2 のような, 現実の状況に近い環境を用いて, カリキュラム学習をした場合としていない場合の結果を比較し, カリキュラム学習の効果の考察を行う.

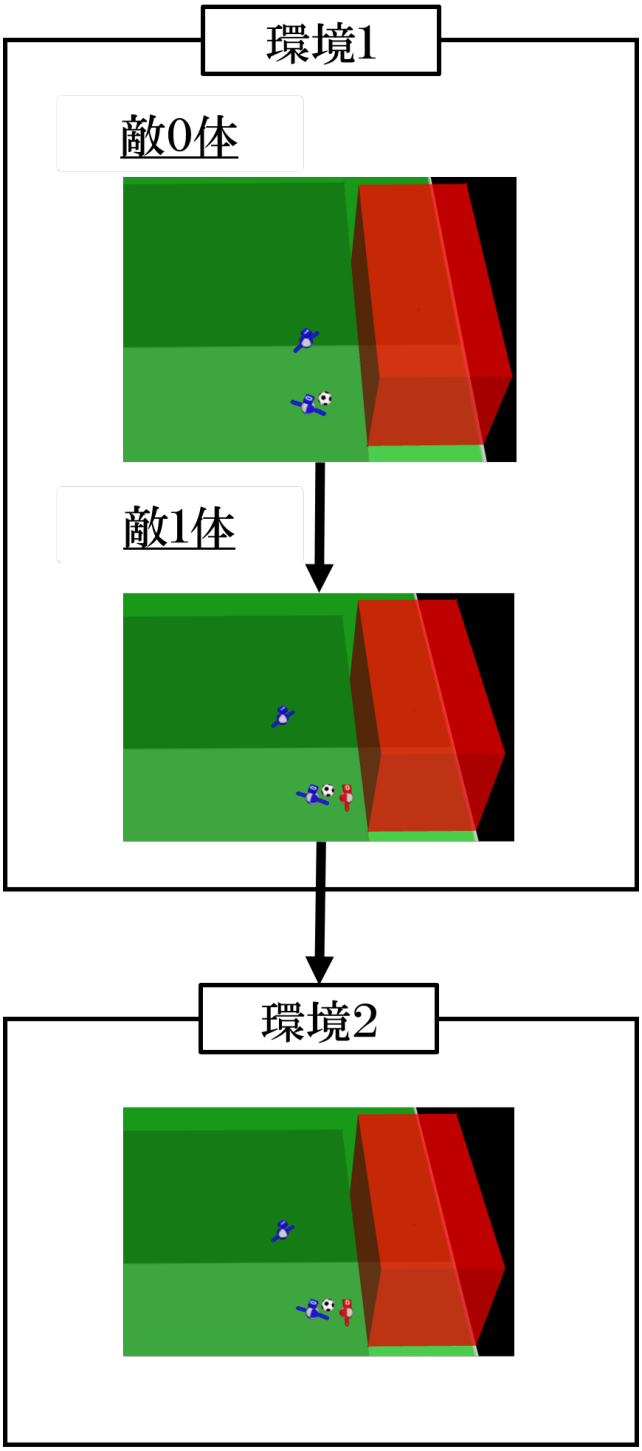


図 3.2　カリキュラム学習　環境 1：条件あり（ 2 体がボールに関与してゴールすれば目標達成） 環境 2：条件なし（ボールをゴールに蹴りこめば目標達成）

3.2.2　Reward Shaping

本研究では, カリキュラム学習に加えて, 段階的に報酬を与える Reward Shaping[14] を用いて学習を行った. Reward Shaping は, 最終目標を達成したときに発生する“外部報酬”以外に“Shaping 報酬”を設計者が与えることで, サッカータスクのような報酬がスパースであるタスクの学習を促す報酬の与え方である. 本研究では, 外部報酬に加えて小さな Shaping 報酬を与えている. 以下にそれぞれの報酬の与え方を示す.

- 外部報酬
  - 2 体のエージェントが触れてゴール +10
- Shaping 報酬
  - 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

本研究では, エージェントがボールに積極的に関わることができるように, 各エージェントが初めてボールに触った時に Shaping 報酬を与えることとした. また, ここでの学習の目標は,2 体のエージェントがボールに関与してゴールすることを学習することである. そのため, 1 体しか触れずにゴールに入った場合は報酬は与えないようにし, エージェントがボールに関わるだけで満足することなく得点を取ることができるように, Shaping 報酬を小さく, 外部報酬を大きく設定している.

3.3　実装アルゴリズム

3.3.1　強化学習

強化学習手法

強化学習はモデルフリー, つまり環境モデルの知識を前提としない状態から学習するため, 様々な行動を試して探索を行い, 最適な行動を選ぶための方策を学習する. 探索では状態の特徴を適切に抽出し, その後の状態を先読みして行動選択する必要がある. この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用いられている.

本研究で取り扱うタスクは, 行動の変数が連続変数であり, 広く知られている Q 学習 [20] を適用するのは難しい. 行動空間が高次元の場合, Q 関数のテーブルも高次元になってしまい, 計算量が大きくなってしまうからである. このような連続値の出力が求められるタスクの場合, 一般的には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[1] 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.

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, 現在の方策によって状態価値を計算する方策評価ステップと, 更新した状態価値関数で価値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 [21].

モデルフリーな方策反復法に対するアプローチとして, 方策勾配法がある. 方策勾配法は, 方策をあるパラメータで表される関数として, そのパラメータを学習させることで, 方策を学習させる方策ベースの手法である. 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 $J(\theta)$ がより大きい値をとるように改善させる. 勾配 $\nabla_\theta J(\theta)$ は以下のように示される [22].

$$\nabla_\theta J(\theta) = \mathbb{E}_\pi[(\nabla_\theta \log \pi(A_t|S_t,\theta))Q_\pi(S_t,A_t)] \tag{3.1}$$

ここではある時間 t において, 状態 S_t のときの行動 A_t を選択した価値関数が $Q_\pi(S_t,A_t)$ とされている. 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ができるようになっている.

REINFORCE アルゴリズム

本研究では, REINFORCE アルゴリズム [23] を用いる. REINFORCE アルゴリズムは, 方策勾配法の一つで, 行動価値関数の簡単な近似として, 割引報酬和 G_t で近似して学習を行うアルゴリズムである. ある時間 t の状態 S_t において, 行動 A_t を選択して将来的に獲得できる報酬

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.

$$G_t = \sum_{k=1}^{T-t} \gamma^{k-1} R_{t+k} \tag{3.2}$$

$$\nabla_\theta J(\theta) \approx \frac{1}{T} \sum_{t=0}^{T-1} \nabla_\theta \log \pi(A_t|S_t, \theta) G_t \tag{3.3}$$

REINFORCE アルゴリズムでは, ある時間 t での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.

3.3.2　本研究で用いたアルゴリズム

本研究では, 伊藤ら [21] が REINFORCE アルゴリズム [23] をもとに作成したプログラムを改変して実験を行った. 図 3.3 に本研究におけるア学習ルゴリズムの全体像を示す. 本実装で用いた REINFORCE アルゴリズムでは, ベースラインとしてアドバンテージ関数を用いる. ここでは, ある時間 t における状態 S_t で行動 A_t を選択したときのアドバンテージ関数 $A^\pi(S_t, A_t)$ を, 以下のように定義する.

$$A^\pi(S_t, A_t) = q_\pi(S_t, A_t) - v_\pi(S_t) \tag{3.4}$$

行動価値関数 $q_\pi(S_t, A_t)$ と価値関数 $v_\pi(S_t)$ の差を表現するアドバンテージ関数 $A^\pi(S_t, A_t)$ は状態 S_t で選択した行動 A_t が, 平均的な行動選択に比べてどれくらいアドバンテージがあるか, つまり有利であるかを表している. このアドバンテージ関数とモンテカルロ近似で期待値を計算し, 行動価値関数 $q_\pi(S_t, A_t)$ を割引報酬和 G_t で近似すると, 勾配 $\nabla_\theta J(\theta)$ は以下のようになる.

$$\nabla_\theta J(\theta) \approx \frac{1}{T} \sum_{t=0}^{T-1} \nabla_\theta \log \pi(A_t|S_t, \theta) A^\pi(S_t, A_t) \tag{3.5}$$

$$A^\pi(S_t, A_t) = G_t - v_\pi(S_t) \tag{3.6}$$

このアルゴリズムでは, 方策パラメータ θ の勾配 $\nabla_\theta J(\theta)$ を用いて, 目的関数 $J(\theta)$ を最大化するように方策を更新する. また, 価値関数 $v_\pi(S_t)$ もニューラルネットなどの関数近似器の出力 $V(S_t)$ によって近似している.

エージェントは状態を観測したあと, ガウスモデルに基づいた確率的方策によって行動を予測する. 連続制御を取り扱う確率的方策 π の一つとして, ガウスモデルによる確率的方策がある. ガウスモデルは, ある時間 t の状態 S_t で平均 $\mu(S_t)$, 共分散行列 $\sum(S_t)$ を持つ K 次元正規分布に沿って K 次元の行動ベクトル A_t をサンプリングする確率モデルである. ガウスモデルは以下の式によって表される.

$$\pi(A_t|S_t, \theta) \propto \frac{\exp\left(-\frac{1}{2}(A_t - \mu(S_t))^T \sum(S_t)^{-1}(A_t - \mu(S_t))\right)}{\sqrt{\det \sum(S_t)}} \tag{3.7}$$

このガウス方策を用いて, (5) の $\log \pi(A_t|S_t, \theta)$ は, 以下のように表される.

$$\log \pi(A_t|S_t, \theta) \propto \sum_{k=1}^{K} \left(\log \sigma_k^2(S_t) + \frac{(A_{tk} - \mu_k(S_t))^2}{\sigma_k^2(S_t)}\right) \tag{3.8}$$

方策としてガウス方策を取り入れることで, REINFORCE アルゴリズムとベースラインを用いたパラメータ更新を行っている.

エージェントが行動を行った後, 観測した状態と報酬を履歴として保存しておき, 終了条件が満たされない場合はステップを継続させる. 本研究では, 最終目標が達成されるまで報酬設

計者によって段階的に与えられる“Shaping 報酬”と最終目標が達成されたときに獲得できる“外部報酬”をエージェントに与えることとしている. ステップごとに与えられる報酬 R_{t+1} は, ステップごとの外部報酬 $R_{extrinsic}$ と Shaping 報酬 $R_{shaping}$ の和で以下の式のように表される.

$$R_{t+1} = R_{extrinsic} + R_{shaping} \tag{3.9}$$

割引報酬和 G_t もステップごとに獲得できる外部報酬 $R_{extrinsic}$ と Shaping 報酬 $R_{shaping}$ から計算される.

また, エピソード終了時, 各ステップの履歴として保存された状態, 行動, 報酬を用いてモデルの更新を行う. 方策の更新は上述の通りであり, 状態価値関数の近似値 $V(S_t)$ の更新は, 以下の式を最小化させるように更新させる.

$$\sum_{t=0}^{T-1}(V(S_t) - G_t)^2 \tag{3.10}$$

価値関数の近似値 $V(S_t)$ が割引報酬和 G_t に近づくようにパラメータの更新を行っている.

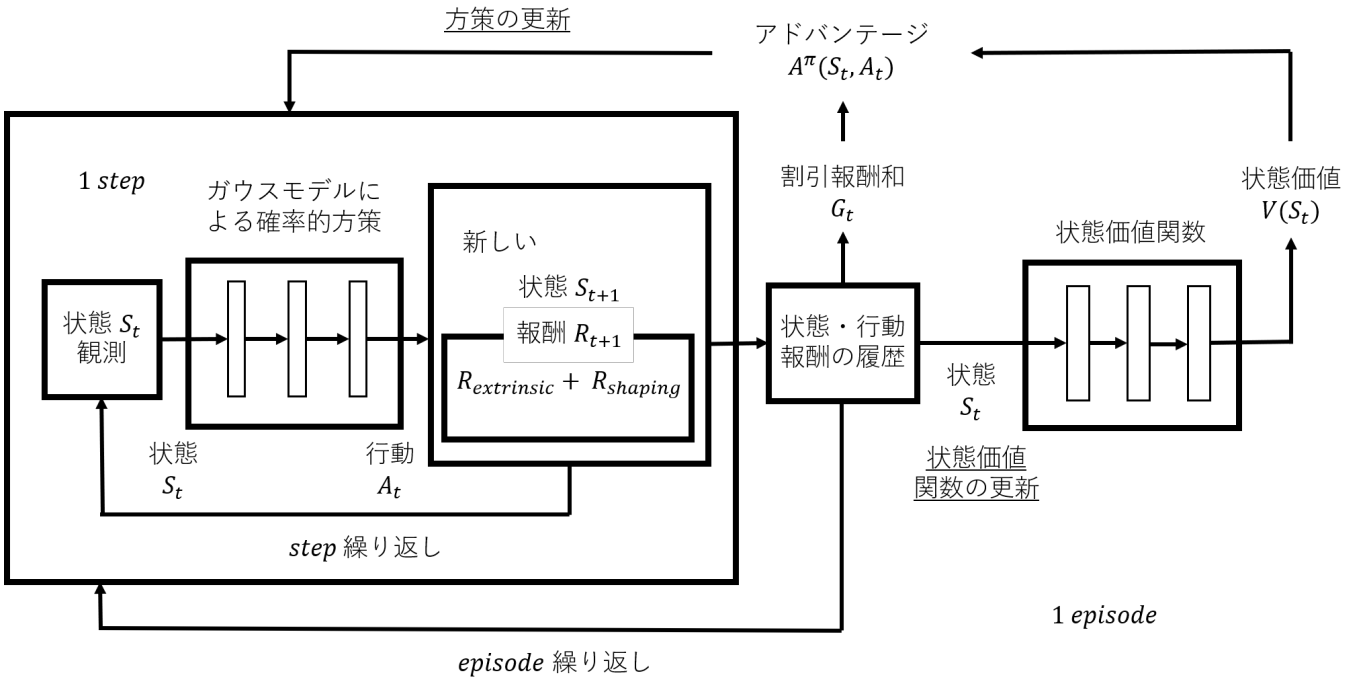


図 3.3　アルゴリズム全体像

第 4 章

実験

4.1 実験準備

本研究では, 物理演算エンジン MuJoCo[9] を用いてサッカータスクでのシミュレーションを行った. サッカータスクは以下の条件でシミュレーションを行った.

状態 位置, 速度, 加速度などのエージェントの情報, エージェントから見たボールの位置, 速度, 角速度, ゴールとコーナーの位置などのピッチの状況, 味方と敵の方向, 位置, 速度を含めた 107 次元の状態空間を持つ.

行動 体を前後を加速させる値, 体を回転させる値, 下向きの力を加えてジャンプさせる値の 3 次元で行動を行う.

エピソード 1 エピソードは最終目標達成時, ボールがピッチ外に出た時, または 10 秒後に終了する.

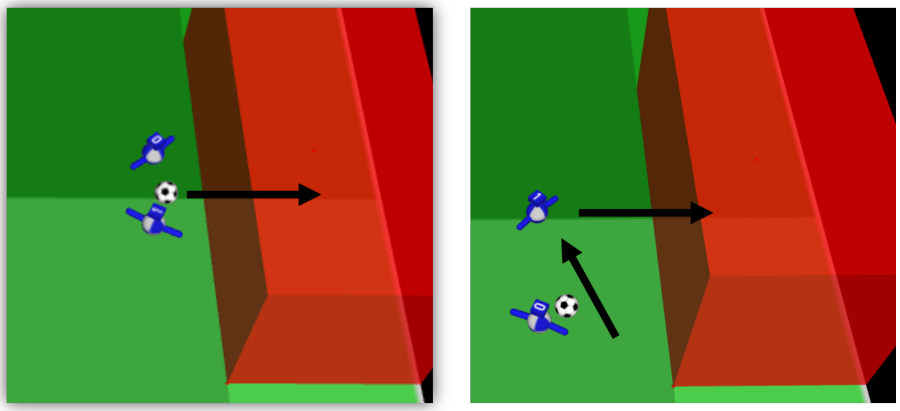
方策は, 入力は 107 次元の状態, 出力は 3 次元の行動で, それぞれ 810, 220, 60 ユニットの隠れ層を持つ 3 層のニューラルネットで実装されている. 価値関数モデルは入力が 81 次元の状態, 出力が 1 次元の行動の価値で, それぞれ 810, 63, 5 ユニットの隠れ層を持つ 3 層の完全結合ニューラルネットワークで実装されている. 学習率は 0.99, 最適化アルゴリズムとして方策の学習に RMSProp を, 価値関数モデルの学習には Adam を用いて学習を行った.

4.2　実験概要

4.2.1　実装手順

本研究における実装の全体像を以下に示す. 実験では, カリキュラム学習における実験の前に, カリキュラム学習を促すための Reward Shaping についても考察を行うため, Reward Shaping における実験も行った.

Reward Shaping における実験では, 図 4.1 の環境を想定して, 実験を行った. 実験 1 ではどちらかがゴールを決めれば目標達成, 実験 2 では 2 体のエージェントが関与して得点を決めることができれば目標達成とする. カリキュラム学習における実験では, 図 4.2 の全体像で実装を行った. まず, 敵 0 体の学習を行い, 徐々に敵を増やして学習を行う. その際, 段階的な学習を行って協調行動を獲得するため, 動かない敵を想定してパス交換の行動獲得をしてから, 動く敵に対する実験を行う.



（1）実験1　　（2）実験 2

図 4.1　Reward Shaping における実験

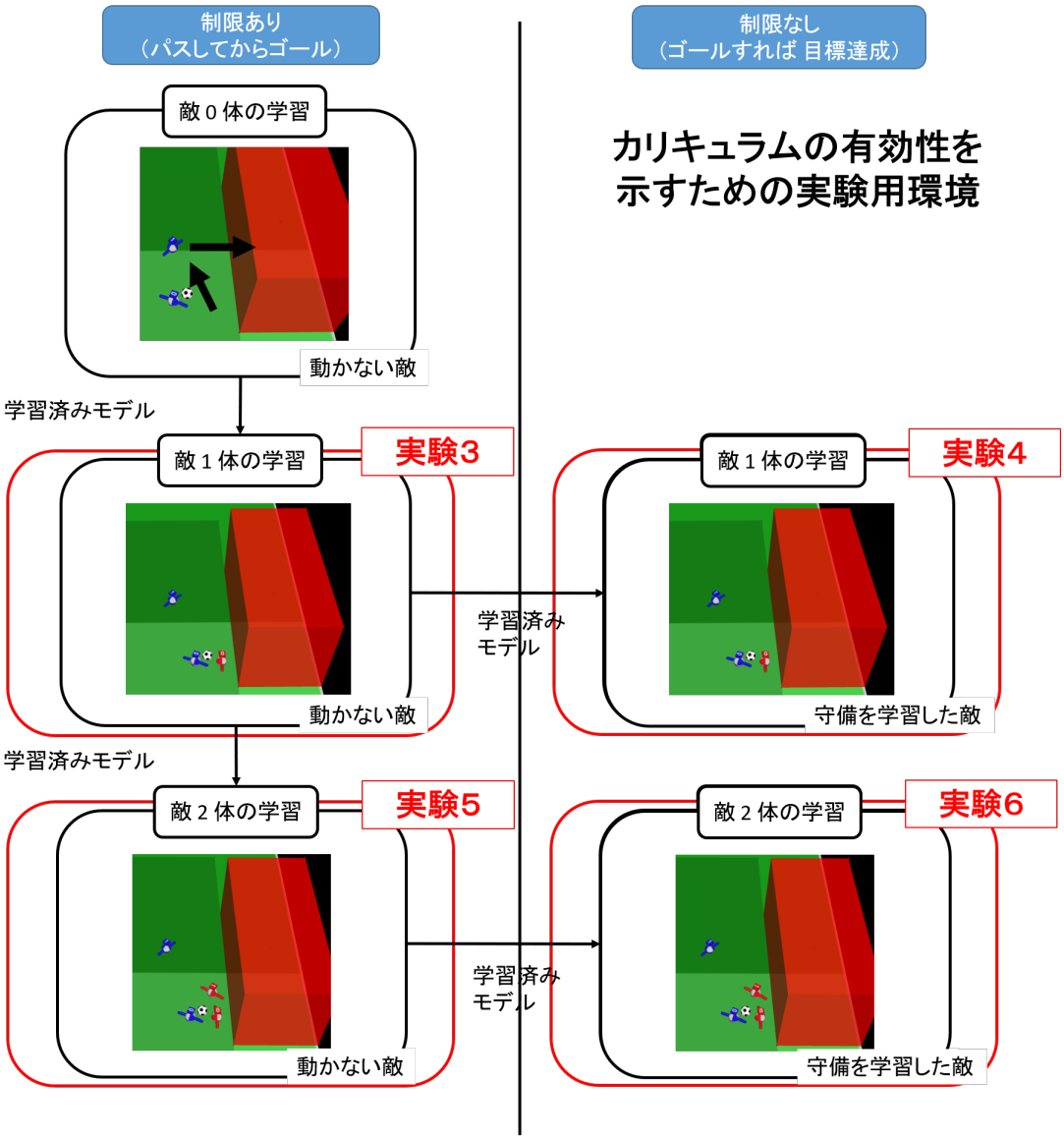


図 4.2　カリキュラム学習における実験

4.2.2　Reward Shaping における実験

Reward Shaping における実験を実験 1,2 で行った. 実験 1,2 では, エージェント間の協調行動を必要とする場合と, そうでない場合で学習を行い, Reward Shaping の効果の考察を行った.

実験 1

実験 1 では, 協調行動を行わない場合, タスクの難易度がどれほどのものであるかを示すために, 図 4.3 のような環境を想定し, 2 体のエージェントのどちらかがボールをゴールに蹴りこめば目標達成という単純な課題を与えて実験を行った. この環境では, 味方の動きに関わらずゴールを決めた時に最終目標達成とおり, Shaping 報酬は与えずに実験を行う.

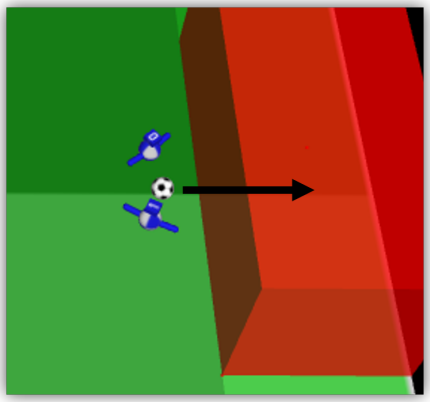


図 4.3　実験 1 の想定環境

実験 2

実験 2 では, 実験 1 を踏まえて, 協調行動の獲得で Reward Shaping をした場合, どれほどの効果があるかを示す. Reward Shaping がサッカータスクでの協調行動にどのような効果をもたらすかを考察するために, Chitnis らによって Mujoco のサッカータスクで行われた研究 [24] に似せた想定環境を作成し, 研究を行った. 実験 2 は Chitnis らの研究と同じ環境を与えており, 2 体のエージェントがボールに関わった後にゴールすれば目標達成とするタスクである. そのため, この環境では片方のエージェントのみの力でゴールを決めたとしても目標達成とはせず, 報酬は与えられない. 人が実際にサッカーの試合をする場面でも, シュートチャンスでパスを選択することで, ゴールの可能性が高くなる場合がある. 実験 2 では, そのような複数のエージェントが協力することで目標達成とすることで, Reward Shaping がマルチエージェントの強化学習にどのような影響を及ぼすか評価し, Shaping ありと Shaping なしで比較を行った. 図 4.4 に実験 2 の想定環境を示す.

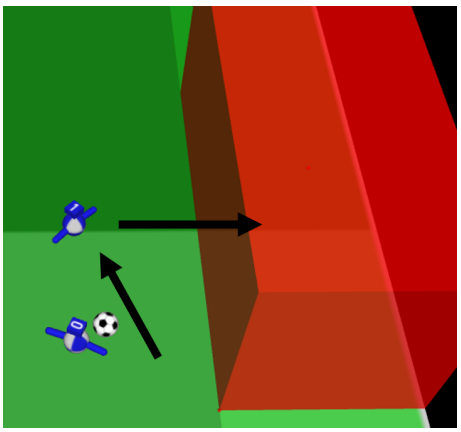


図 4.4　実験 2 の想定環境

4.2.3　カリキュラム学習における実験（敵 1 体）

実験 3,4 では, Chitnis らの研究 [24] の想定環境に守備エージェントを追加した, 図 4.2 のようなシュートチャンスで実験を行い, カリキュラム学習における協調行動の獲得を試みた.

実験 3

実験 3 では, 図 4.1 で示された実験 2 の想定環境, つまり敵がいない環境で学習された学習済みモデルを用いて, 守備エージェント 1 体を加えて行った学習における結果を示す. 実験 3 では, エージェントは図 4.5 のような初期配置で固定された状態で開始し, 守備エージェントは動かない障害物として学習を行う. カリキュラム学習では, 協調行動の獲得を促すことを目的とするため, 実験 3 の想定環境では, 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, この環境では片方のエージェントのみの力でゴールを決めたとしても目標達成とはせず, 報酬は与えられない. カリキュラム学習を用いたものを Curriculum, 用いなかったものを No Curriculum として学習の比較を行った.

実験 4

実験 4 では, 実験 2,3 で協調行動を学習させた学習済みモデルを用いて実験を行い, カリキュラム学習の効果を検証した. 実験 4 では, エージェントは図 4.5 のような初期配置で固定された状態で開始し, 守備エージェントは相手に簡単にシュートを打たせないように守備を学習した状態で, 攻撃側のエージェントの学習を行う. また, 実験 4 での目標達成は制限をせず, ボールをゴールに入れたら目標達成とする. カリキュラム学習を用いたものを Curriculum, 用いなかったものを No Curriculum として学習の比較を行った.

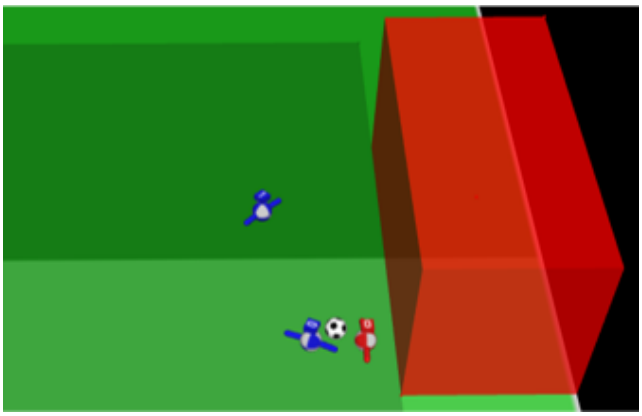


図 4.5　実験 3,4 の想定環境（実験 3：動かない敵　実験 4：動く敵）

4.2.4　カリキュラム学習における実験（敵 2 体）

実験 5

実験 5,6 でもカリキュラム学習における実験を行う. 実験 5 では, 実験 3 の環境に守備エージェントを 1 体追加して実験を行った. 守備エージェントが増えると, 攻撃側のエージェントは考慮すべき情報量が多くなるため, 学習が難しくなる. 実験 5 の想定環境を図 4.6 に示す. 実験 3 と同じように, 実験 5 では 2 体のエージェントがボールに関わった後にゴールすることができれば目標達成としている. そのため, この環境では片方のエージェントのみの力でゴールを決めたとしても目標達成とはせず, 報酬は与えられない. 実験 5 では, 実験 2,3 で学習された学習済みモデルを使って, 学習の効果をみる. カリキュラム学習を用いたものを Curriculum, 用いなかったものを No Curriculum として学習の比較を行った.

実験 6

カリキュラム学習の効果を検証するため, 実験 2,3,5 で学習された学習済みモデルを用いて, 以下の図の想定環境で実験を行った. 実験 6 では目標達成に制限はなく, ボールをゴールに入れたら目標達成とする. 実験 6 では, エージェントは図 4.6 のような初期配置で固定された状態で開始し, 守備エージェントは相手に簡単にシュートを打たせないように守備を学習した状態で, 攻撃側のエージェントの学習を行う. カリキュラム学習を用いたものを Curriculum, 用いなかったものを No Curriculum として学習の比較を行った.

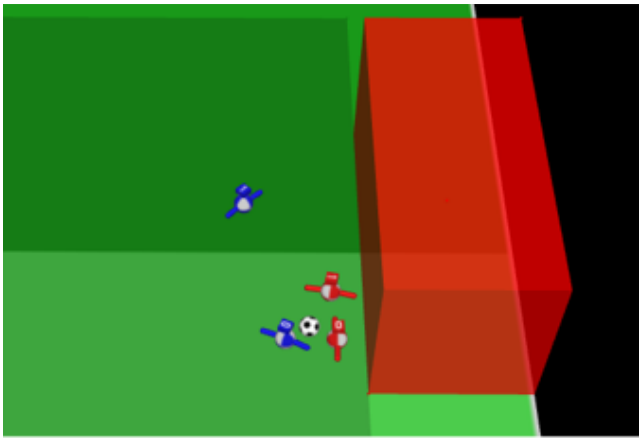


図 4.6　実験 5,6 の想定環境（実験 5：動かない敵　実験 6：動く敵）

4.3　実験結果

4.3.1　Reward Shaping における実験

実験 1

エージェント間の協調行動が必要のない実験 1 の環境において, 平均目標達成率で学習効果の評価を行った. その結果を図 4.7 に示す.

この実験では, 最終目標を達成したときのみ外部報酬を獲得することができる. 通常のサッカーの試合であれば, 報酬がスパースで外部報酬を獲得することが難しくなるが, 実験 1 の想定環境では, ゴールまでの距離が近く, ボールを触れるだけでもゴールになるため, 容易に目標を達成することができる. 図 4.7 を見ると, 最初から 5 割以上の目標達成率を維持し続け, およそ 10000 エピソードあたりから学習が加速し, およそ 30000 エピソードでは 7 割で目標達成することができている. このような容易なタスクであれば, 低レベルからの行動獲得でも比較的早く学習を進めることができる. 実験 1 の環境は, 実験 2 の環境と異なり, 協調行動の必要がなく, 味方の状態を考慮する必要が少ないため, 学習を容易に進めることができたとみれる.

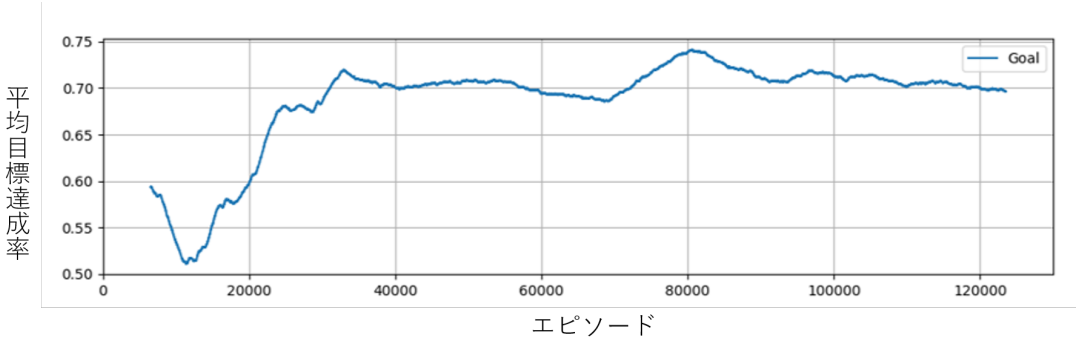


図 4.7　実験 1 の実験結果

実験 2

エージェント間の協力が必要な実験 2 について, 平均目標達成率で学習効果を評価した. 図 4.8 は, Reward Shaping を行った場合と行わなかった場合の結果を比較したものである. 学習の結果, Shaping ありの場合の平均目標達成率が, Shaping なしの場合に比べて高くなっていることがわかる.

これらの比較から, Shaping 報酬が効率よく与えられ, 学習を妨げることなく学習を促していることがわかり, 2 体のエージェントの協調行動の獲得に良い影響を与えることがわかる. 逆に, Shaping 報酬なしの結果をみると, エージェントが相互作用を必要とする場合には, 相互作用を必要としない場合に比べて, 低レベルからの行動獲得が難しいことを示している. 複雑な動きがあらかじめ実装されていない場合, エージェントの行動の選択肢が多く, 探索が困難になる. 実験 2 の想定環境では, パスを出すエージェントは味方の位置を把握してそこに向けてパスを出し, 次にパスをもらうエージェントは, ボールをコントロールしてどのようにゴールを決めるかを学習しなければならない. これらの一連の動作が全て上手くいったときのみに報酬が与えられるため, これらの行動を 2 体のエージェントが少ない報酬で学習することは難しい.

図 4.9-4.11 に, 学習に Shaping 報酬を用いた場合の, 平均獲得報酬, 平均獲得 Shaping 報酬, 平均外部獲得報酬を示す. 平均獲得 Shaping 報酬をみると, 最終的に報酬は 2.0 に収束し, 2 体のエージェントがボール関与することを学習することを Reward Shaping によって促すことができているとみることができる. エージェントが初めてボールに関与したときに 1.0 の Shaping 報酬を獲得するため, 学習が進むにつれてほぼ全てのエピソードで, 2 体のエージェントがボールに関与していることになる. はじめに 2 体のエージェントがボールに関わりやすい, つまり 1 体目のエージェントが 2 体目のエージェントに向かってボールを蹴りやすくなり, シュートチャンスを増やすことが, より高い目標達成率の獲得を促すことができていると考えられる.

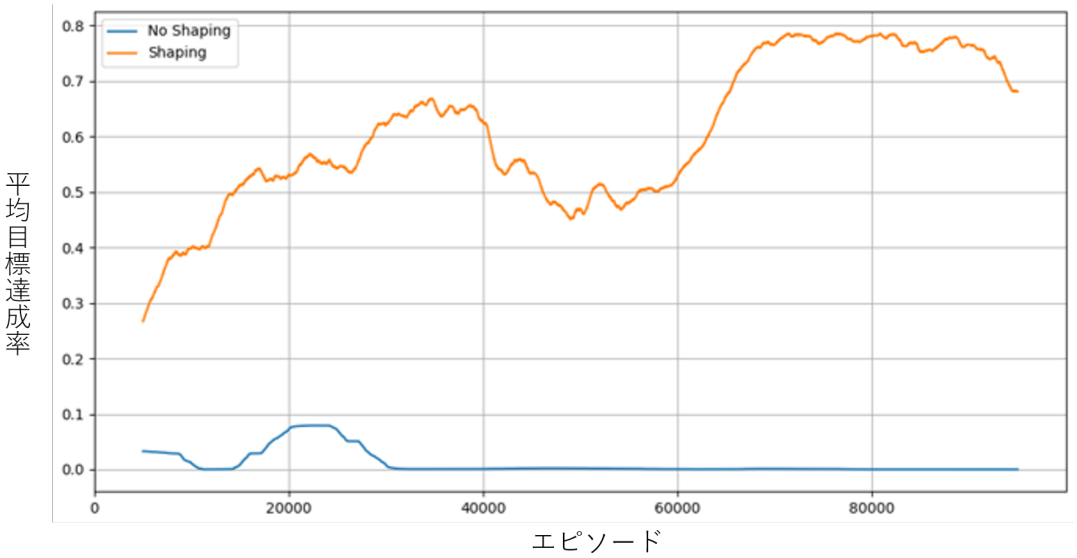


図 4.8　実験 2 の実験結果

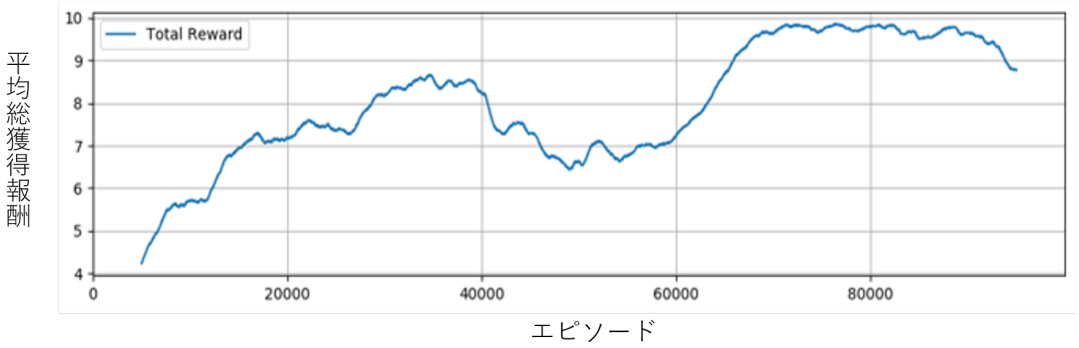


図 4.9　実験 2 の平均総獲得報酬

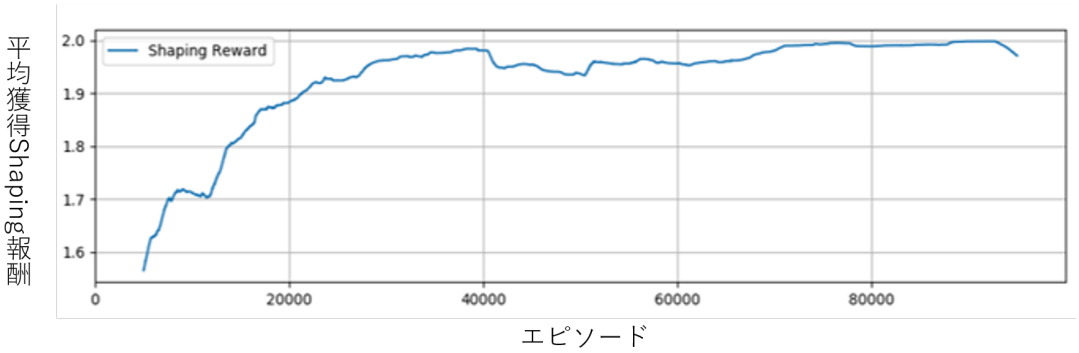


図 4.10　実験 2 の平均 Shaping 獲得報酬

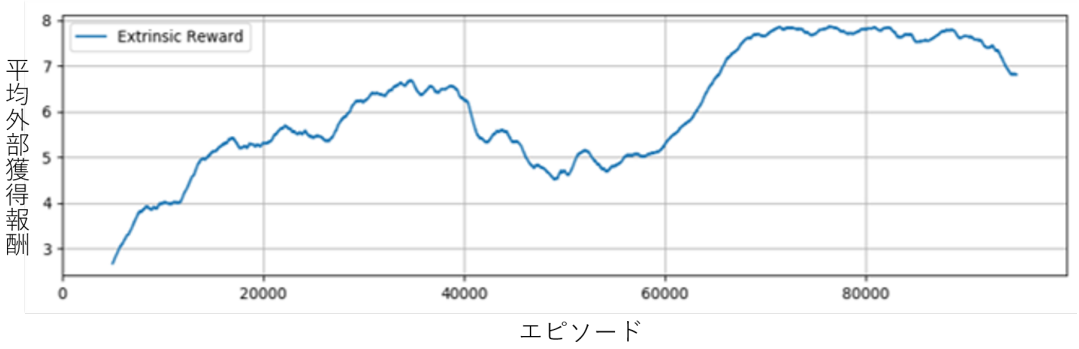


図 4.11　実験 2 の平均外部獲得報酬

4.3.2　カリキュラム学習における実験（敵 1 体）

実験 3

カリキュラム学習された学習済みモデルを用いた Curriculum, 事前学習せずに学習させた No Curriculum を使って, 図 4.5 の敵 1 体の環境で比較実験を行った. 学習済みモデルは敵 0 体のタスクで 10000 エピソード学習させたものを使用した. 図 4.12 にそれぞれの平均目標達成率を示す.

Curriculum は 5 割程度まで目標達成率が到達したのに対して, No Curriculum は学習が進んでいないことがわかる. Curriculum の学習済みモデルは敵がいない状態で味方とのパス交換を学習しており, 敵がいる場合の状態にも対応できていることがわかる. 実験 3 では, 敵がパス交換にあまり干渉しないことから Curriculum は効率よく学習できている. 逆に No Curriculum では, 例えば片方のエージェントが少し触るだけの場合, 敵のエージェントをかわすことができず, ゴールをすることができないことがある. このように事前学習をしない場合, ゴールできる機会, つまり報酬が得られる機会が少なくなり, 学習が進まなくなってしまう.

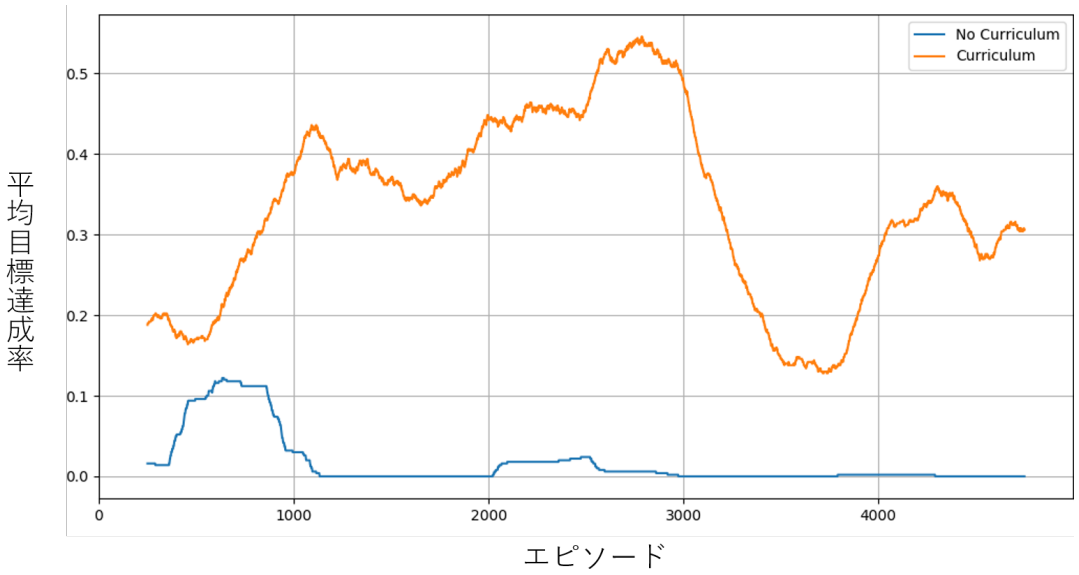


図 4.12　実験 3 の実験結果

実験 4

カリキュラム学習された学習済みモデルを用いた Curriculum, 事前学習せずに学習させた No Curriculum, 比較用として図 4.4 の敵 0 体の状態で行った学習済みモデルを用いた Light Curriculum を使って, 比較実験を行った. Curriculum は敵 0 体のタスク, 敵 1 体のタスクそれぞれで 10000 エピソード学習させた学習済みモデルを使用した. 図 4.13 にそれぞれの平均目標達成率を示す.

それぞれの目標達成率を比較すると, Curriculum は序盤から約 5 割の目標達成率であるのに比べて, No Curriculum は 4 割を超えるまでにおよそ 2000 エピソードかかっており, Light Curriculum は全く学習が進んでいないことがわかる. Curriculum は 1000 エピソードまでにも目標達成率を約 4 割を達成しており, カリキュラム学習による影響が出ていることがわかる. 敵がいない状態や敵が動かない状態での学習によって, シュートパターンを学習させておくことは, ボールを奪いに来る守備エージェントがいる場合にゴールを奪う手掛かりとなり, 環境 2 においてもゴールを決めやすくなっていることがわかる. No Curriculum は, ゴールできるときとできないときがまばらなため, 探索に時間がかかり, 5000 エピソード前後では目標達成率が 1 割程度である. 協調行動を必要とする複雑なタスクでは, このように学習が難しくなるため, タスクに適したアプローチが必要になる. また, Light Curriculum は全く学習が進んでいないことがわかる. 敵がいない状況での学習のみの場合, 動く敵に対応することができず, 事前学習で学習されたことが逆に学習の妨げとなってしまっていたと考えられる.

これらの実験結果を比較すると, カリキュラム学習を行った場合でも, 学習を効率的に進められる場合とそうでない場合があることがわかる. Curriculum のように敵のいない状況での学習と動かない敵を交えた学習を行った場合, 敵を交わしてゴールすることを学習しているが, Light Curriculum では敵を交わす学習ができていない. Light Curriculum のように事前学習を行うと, 学習に悪影響を及ぼしてしまうことがわかる. このように, カリキュラム学習で行動獲得をする場合, カリキュラムとカリキュラムを構成するタスクは解決するタスクに適したものを考える必要がある. そのため, カリキュラム学習を効果的に行うためには, 解決したいタスクへのドメイン知識を持った設計者によって適切なカリキュラム構築をすることが求められる.

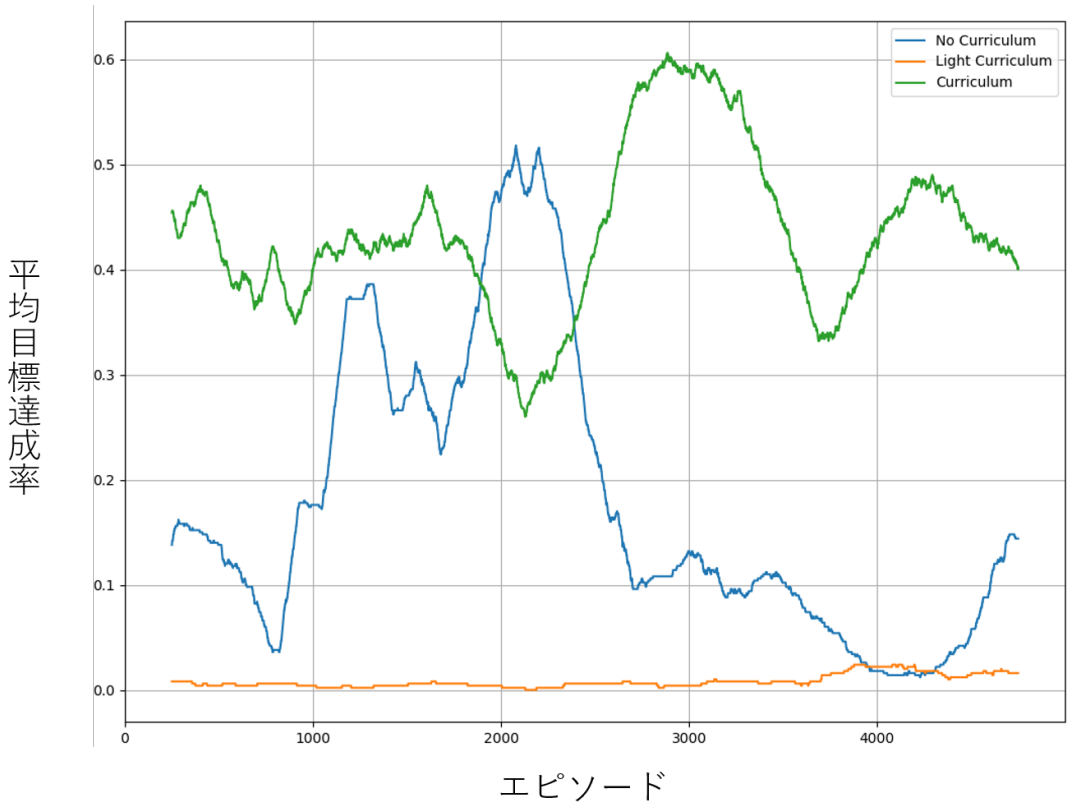


図 4.13　実験 4 の実験結果

4.3.3　カリキュラム学習における実験（敵 2 体)

実験 5

カリキュラム学習された学習済みモデルを用いた Curriculum, 事前学習をしなかった No Curriculum で図 4.6 のような環境で比較実験を行った. カリキュラム学習は敵 0 体のタスク, 敵 1 体のタスクそれぞれで 10000 エピソード学習させた. 図 4.14 にそれぞれの平均目標達成率を示す.

Curriculum は No Curriculum に比べて平均目標達成率が高いことがわかる. 1000 エピソードまではどちらも 1 割程度の成功率だったが, 以降は Curriculum の学習が進んでいるのがわかる. 特に 2500 エピソード学習させると, 6 割近く目標達成していることがわかる. 敵がいない場合のタスク, 敵が 1 体いた場合のタスクを学習させることで, エージェントのいないところでパスを通すというスキルを学習させることができ, 敵が増えた場合でも目標に早く到達することができた. 一方で No Curriculum は報酬を得られる機会が少なく, 学習が進んでいない. Shaping 報酬を与えたとしても, 障害物が多くある場合では学習が進みにくいことがわかる. Curriculum と No Curriculum の比較から, ドメイン知識を持った設計者によるカリキュラム学習を行うことによってエージェントが効率的に学習を行うことができたとみることができる.

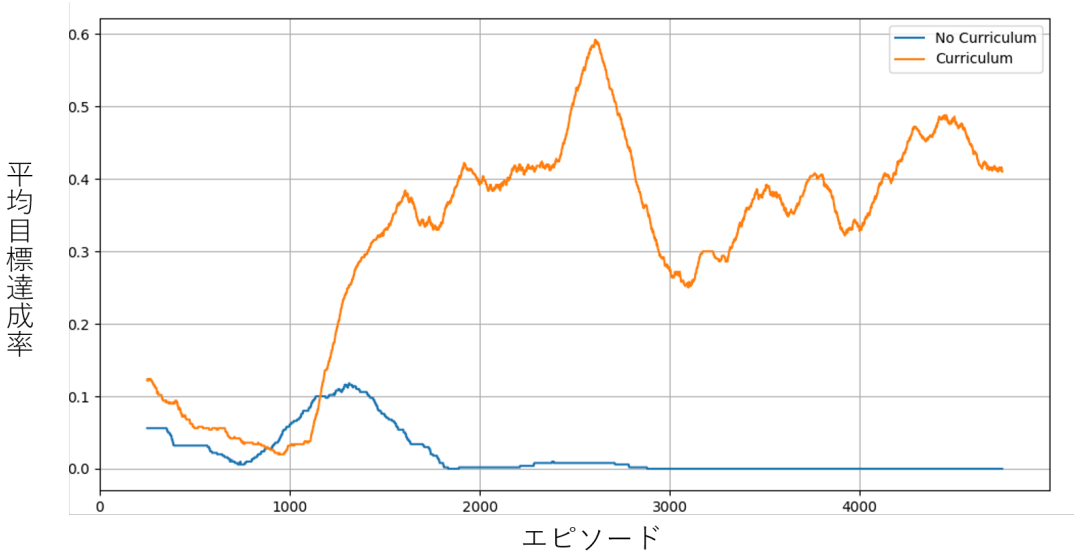


図 4.14　実験 5 の実験結果

実験 6

カリキュラム学習された学習済みモデルを用いた Curriculum, 事前学習せずに学習させた No Curriculum を使って, 比較実験を行った. Curriculum は敵 0 体のタスク, 敵 1 体のタスクそれぞれで 10000 エピソード学習させ, 敵 2 体のタスクで 20000 エピソード学習させた学習済みモデルを使用した. また, 外部報酬を 10 ではなく, 100 にして外部報酬の影響を大きくして学習を促している.

図 4.15 にそれぞれの平均目標達成率を示す. 5000 エピソードまで学習させると, No Curriculum が Curriculum を上回っていることがわかる. 2 体の守備を学習した守備エージェントが配置されている場合, 1 体はゴールを守り, 1 体はパスコースを防ぐような行動を取ることがある. 本研究で用いたカリキュラム学習で学習させたパスコースは, パスが防がれてしまい, カリキュラム学習が目標達成に悪影響を及ぼしてしまっている. Curriculum がカリキュラム学習によって決められたパスを学習しても, そのパスコースを防がれてしまった場合に対応できないため, 本研究でのカリキュラム学習では守備エージェントを増やしたり, 守備エージェントの配置によっては学習が進まなくなってしまうことがわかる.

図 4.16 に 140000 エピソードまで学習させたものを示す. 5000 エピソードまでの学習では, 目標達成率はどちらもほぼ 1 割以下であったため, 長い時間学習させたモデルの比較を行った. 長い時間学習させると, 報酬を得る機会が増え, 学習が進んでいくことがわかる. 特に 80000 エピソードで Curriculum の目標達成率が 2 割到達, 120000 エピソードで No Curriculum の目標達成率が 2.5 割到達している. 守備を学習した 2 体の守備エージェントがいる場合, 学習が難しく, カリキュラム学習をした場合としない場合で差を見出すことができず, カリキュラムで学習したこととは別の手段でゴールを目指していると考えられる. 本研究では,1 つのパスコースに限定させてカリキュラム学習をしたが, 学習の段階でいくつかのスキルを学習させることや, 少しずらしたパスコースの学習など, 攻撃エージェントが守備エージェントの行動に対して臨機応変に対応できるようなカリキュラムを構成することが求められる.

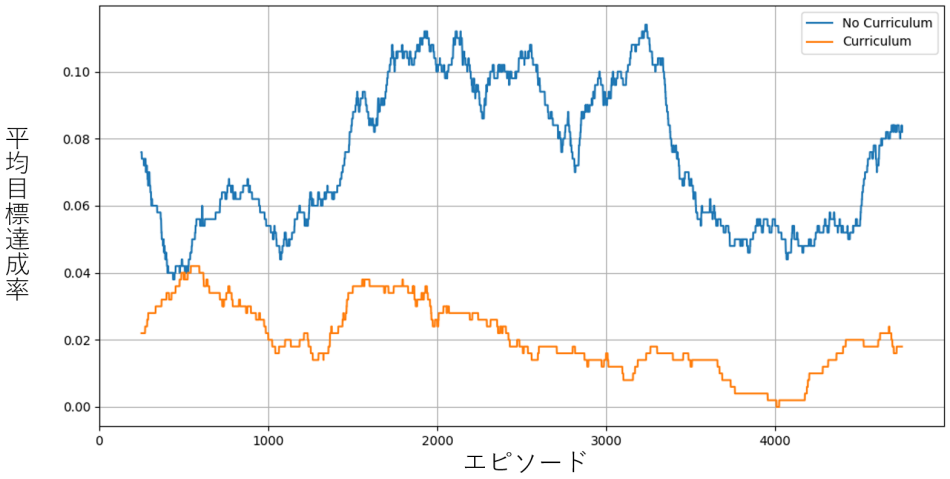


図 4.15　実験 6 の実験結果

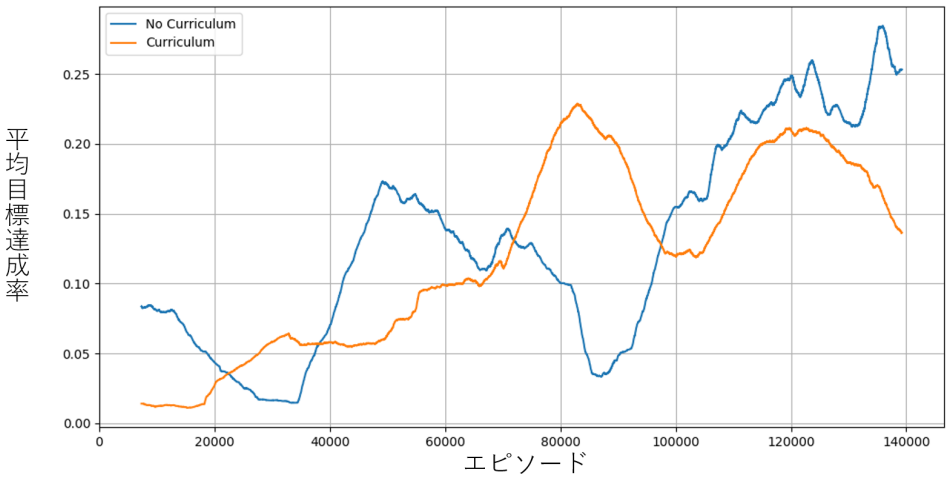


図 4.16　実験 6 の実験結果（140000 エピソード）

第 5 章

考察

5.1　カリキュラム学習における協調行動獲得

本研究では, 動かない相手をコーンとして見立ててカリキュラム学習を行うことが, 適切なカリキュラムを組むことができれば, サッカータスクの強化学習においても有用であることを示した. サッカーでは, ボールを奪われることなくゴールまでボールを運んでいく必要があり, 予測することが難しい相手の行動を考慮した学習が必要になる. 敵がいる場合といない場合では想定する環境が大きく変わってくるため, コーンのように動かない敵を配置してサッカーの練習のようにエージェントを学習させることはカリキュラム次第で, 学習に好影響を与えることができる.

実験では攻撃エージェント 2 体と守備エージェントのシュートチャンスに限定しているが, 本研究の想定環境を拡張することで様々な攻撃パターンを学習できる. サッカーでは, 様々なパターン練習を行い, 実際の試合に備える. 例えば, サイドから中央に向けてボールを入れてチャンスを作るセンタリングを想定する場合, 相手と相手の間に攻撃エージェントが入り込んでボールに合わせることを学習できる. 他にも, スルーパスを出すパターンを想定する場合, コーンを置いて相手の背後に向けてパスを出すことや, ゴールキーパーの立ち位置を想定してシュート練習をすることで, 正確なコースへのシュートを学習することもできる. このように, 多様な協調行動のスキルを獲得するために, 障害物を配置してカリキュラム学習を行うことは効果的であると考えられる. そのためには, それぞれのパターンに合うカリキュラムを考えることが必要であり, ドメイン知識を持つ設計者によるカリキュラム構築が必要になる.

5.2　カリキュラム学習におけるタスク設計

実験 4 で, 敵がいない場合のみ事前学習したとき, 学習に悪影響を及ぼしたこと, 実験 6 ではカリキュラム学習自体の効果が薄かったことから, カリキュラム学習では設計者による適切なタスク設計が重要になることがわかる. 固定された少ない状況のパターンをカリキュラム学習をすると, より複雑な状態が表れた時に, 学習に悪影響を及ぼしてしまう. これまでの研究の中で, パラメータをランダムに設定して, 多くの状況のタスクを自動生成することを通して, エージェントが様々な状況に対応できるように学習させる研究が行われてきた [16]. このように, 多くのタスクを生成して, 様々な環境でも対応できるエージェントをカリキュラム学習させていくことが必要になると考えられる. 本研究では, まだこれらに取り組むことができておらず, より複雑な環境でのカリキュラム学習の適用は今後の課題としている.

5.3　低レベルからの行動獲得

また, 本研究では, 低レベルからの行動獲得も可能であることを示した. これまでのサッカータスクにおけるカリキュラム学習の研究 [16][18] では, パスやシュートなどの基本的なスキルはあらかじめ設計されて学習が行われていた. 本研究では, 3 次元の値で表現された行動で学習を行っており, 体を前後を加速させる値, 体を回転させる値, 下向きの力を加えてジャンプさせる値で学習を行った. 実験より, カリキュラム学習を用いて, 低レベルからでもシュートチャンスでの協調行動を獲得できることがみてとれる. 低レベルから協調行動を獲得できることで, エージェントごとのスキルの設計の手間を省くことができるという利点がある. サッカータスクのスキルは, RoboCup などで手作業で長年にわたり設計されてきた. しかし, これまで洗練されてきたエージェントより優れたエージェントのスキルの作成には, かなりの時間を要する. その点, これまでに作成されたサッカーエージェントを相手に, 低レベルからエージェントが学習を行うことができれば, 設計技術を持たない設計者でも容易に優れたエージェントの作成を可能にすることができる.

第 6 章

おわりに

本研究では, サッカータスクでの協調行動の獲得を目指し, カリキュラム学習を用いた強化学習の研究を行った. 本稿では, シュートチャンスでのゴール率を高めるため, 段階的に難しいタスクを学習させるカリキュラム学習を強化学習に適用させ, Reward Shaping を加えて学習を行った. まずは敵のいない状態での学習, 次に動かない敵を配置しての学習, 最後に守備を学習したエージェントを配置しての学習を行い, それぞれの学習で学習された学習済みモデルを使用することでカリキュラム学習を行った. 実験では, 協調行動が必要なシュートチャンスにおいて, 協調行動をカリキュラム学習した場合は, カリキュラム学習しなかった場合に比べて目標達成率が高くなることが示された. 敵のいない状態での学習や, 敵の動かない状態でのカリキュラム学習が最終目標達成を促すことができていると考えられる. また, 設計者による不適切なカリキュラム学習は学習に悪影響を及ぼしてしまうこともわかり, ドメイン知識を持つ設計者による適切なカリキュラム作成が必要であることもわかった. 今後は, 環境設定においてエージェント数を増やしたり, ゴールとの距離を離すなど複雑なタスクでのカリキュラム学習の検討や, エージェントごとに異なる方策を用いた場合の学習の検討を行い, 実際の試合に近い形のタスクでの学習を目指したい.

謝辞

本研究を進めるにあたり，ご指導およびご協力いただいた皆様へ心から感謝の気持ちと御礼を申し上げたく, ここに感謝の辞を述べさせていただきます．本研究を遂行するにあたり, 毎週のゼミにてご指導いただいた大須賀昭彦教授, 田原康之准教授, 清雄一准教授に深く感謝申し上げます. 特に, 3 人の先生方が私のためにゼミの時間をとっていただいたことは印象深く, 大変ご迷惑をおかけしたことと共に, 多くの成長の場を与えてくださいました. また, 研究や体調管理のことなど毎週のゼミで気にかけていただいた折原良平客員教授に深く感謝申し上げます. 研究が行き詰まったときに支えてくださり, 心の支えであったとともに, 論文の細かい修正や発表準備を丁寧に指導してくださり, 多くの時間を割いてくださったことに心から感謝いたします. また, 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国立情報学研究所／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いたします.

そして, 卒業研究, 修士研究やコロナ禍での就職活動をともに乗り越えた大須賀・田原・清研究室同期のメンバーの皆様に心から感謝いたします. 3 年間で自分の成長を垣間見えることができたのも, 同期に励まされ, 日頃から努力を続ける同期ゆえでありました. 最後に研究生活を含め, 様々な面でお世話になりました大須賀・田原・清研究室の皆様に, この場を借りて感謝の意を表します.

参考文献

[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis, “Human-level control through deep reinforcement learning”, Nature, 7540, 529-533, 518, feb 2015.

[2] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis, “Mastering the game of go with deep neural networks and tree search”, Nature 529, 7587 (January 2016), 484–489, 2016.

[3] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra, “Continuous control with deep reinforcement learning.”, 4th International Conference on Learning Representations (ICLR ’ 15), 10 pages, 2015.

[4] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. “Deterministic policy gradient algorithms”, In Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (ICML’14), 387–395, 2014.

[5] Minne Li, Zhiwei Qin, Yan Jiao, Yaodong Yang, Jun Wang, Chenxi Wang, Guobin Wu, and Jieping Ye, “Efficient Ridesharing Order Dispatching with Mean Field Multi-Agent Reinforcement Learning”, In The World Wide Web Conference (WWW ’19), Association for Computing Machinery, New York, NY, USA, 983–994, 2019.

[6] Prabuchandran K.J., Hemanth Kumar A.N and S. Bhatnagar., “Multi-agent reinforcement learning for traffic signal control”, 17th International IEEE Conference on Intelligent Transportation Systems (ITSC ‘14), IEEE, 2529-2534. 2014.

[7] Kitano Hiroaki, Asada Minoru, Kuniyoshi Yasuo, Noda Itsuki and Osawa Eiichi, “RoboCup:

The Robot World Cup Initiative”, Proceedings of the First International Conference on Autonomous Agents, 340-347, 8, Marina del Rey, California, USA, 1997.

[8] Karol Kurach, Anton Raichuk, Piotr Stanczyk, Michal Zajac, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet and Sylvain Gelly,“Google Research Football: ”A” Novel Reinforcement Learning Environment”, CoRR, 2019.

[9] Emanuel Todorov, Tom Erez, and Yuval Tassa,“MuJoCo: A physics engine for model-based control”, 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 5026-5033, 2012.

[10] Siqi Liu, Guy Lever, Josh Merel, Saran Tunyasuvunakool, Nicolas Heess and Thore Graepel,“Emergent Coordination Through Competition”, CoRR, 2019.

[11] Sutton, Richard S. and Barto, Andrew G., Introduction to Reinforcement Learning, MIT Press, 1998.

[12] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Weston, Jason, “Curriculum Learning”, Proceedings of the 26th Annual International Conference on Machine Learning, 8, 41-48, Montreal, Quebec, Canada, 2009.

[13] Peter Stone, Richard S. Sutton and Gregory Kuhlmann, “Reinforcement Learning for RoboCup Soccer Keepaway”, Adaptive Behavior, 13, 3, 165-188, 2005.

[14] Andrew Y. Ng, Harada Daishi, and Stuart J. Russell, “Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping”, Proceedings of the Sixteenth International Conference on Machine Learning, San Francisco, CA, USA, 278-287, 10, 1999.

[15] Sam Devlin, Marek Grzeundefined, and Daniel Kudenko, “Multi-Agent, Reward Shaping for RoboCup KeepAway”, The 10th International Conference on Autonomous Agents and Multiagent Systems, Volume 3, 1227-1228, 2, Taipei, Taiwan, 2011.

[16] Sanmit Narvekar, Jivko Sinapov, Matteo Leonetti, and Peter Stone, “Source Task Creation for Curriculum Learning”, International Foundation for Autonomous Agents and Multiagent Systems, 566-574, 9, Singapore, Singapore, 2016.

[17] Shivaram Kalyanakrishnan, Yaxin Liu, and Peter Stone, “Half Field Offense in RoboCup Soccer: A Multiagent Reinforcement Learning Case Study”, 72-85, 2006.

[18] Felipe Leno Da Silva, and Anna Helena Reali Costa, “Object-Oriented Curriculum Generation for Reinforcement Learning.”, In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS ’18), Richland, SC, 1026–1034, 2018.

[19] 野田五十樹, “ロボットにおける機械学習の課題と動向”, 情報処理 44.11, 1145-1151, 2003.

[20] Watkins, Christopher John Cornish Hellaby, Learning from delayed rewards, King’s College,

Cambridge, 1989.

[21] 伊藤 多一, 今津 義充, 須藤 広大, 仁ノ平 将人, 川 悠介, 酒井 裕企, 魏 崇哲, 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.

[22] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour,“Policy gradient methods for reinforcement learning with function approximation”, Advances in neural information processing systems, 1057-1063, 2000.

[23] Ronald J Williams,,“Simple statistical gradient-following algorithms for connectionist reinforcement learning”, Machine learning, 8, 3-4, 229–256, 1992.

[24] Rohan Chitnis, Shubham Tulsiani, Saurabh Gupta, and Abhinav Gupta,“Intrinsic Motivation for Encouraging Synergistic Behavior”, arXiv, 2020.

研究業績

論文誌

1. 阿部宇志, 折原良平, 清雄一, 田原康之, 大須賀昭彦: サッカータスクでの協調行動におけるカリキュラム学習を用いた行動獲得, 人工知能学会特集論文「エージェント技術とその応用 2021」, 査読審査中

国際会議

1. Takashi Abe, Ryohei Orihara, Yuichi Sei, Yasuyuki Tahara, Akihiko Ohsuga: Acquisition of Cooperative Behavior in a Soccer Task Using Reward Shaping, International Conference on Innovation in Artificial Intelligence (ICIAI 2021), 2021.3（発表予定）

研究会

1. 阿部宇志, 折原良平, 清雄一, 田原康之, 大須賀昭彦: 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, SMASH20 SUMMER SYMPOSIUM (2020.9)
2. 阿部宇志, 折原良平, 清雄一, 田原康之, 大須賀昭彦: サッカータスクの協調行動獲得におけるカリキュラム学習を用いた強化学習, 電子情報通信学会, 人工知能と知識処理研究会 (2021.2)