点対称の像を用いた追跡問題の Q 学習の高速化
Using Point-Symmetric Images to Accelerate Q-Learning for Hunting Problems

飯岡　徹人
Tetsuto Iioka
法政大学情報科学部ディジタルメディア学科
E-mail: tetsuto.iioka.7v@stu.hosei.ac.jp

Abstract

Recently, machine learning attracts attention in the field of artificial intelligence. Reinforcement learning, which is a kind of machine learning and includes Q-learning, is good at solving problems whose optimal solutions are unclear. We propose a method for accelerating Q-learning for hunting problems by using point-symmetric images. In a hunting problem, a hunter agent chases a prey agent in a two-dimensional field. In our method, in addition to the pray agent that the hunter agent chases, another pray agent moving at the point-symmetric location to the original pray agent is also used for Q-learning. Our method is a variant of previous work that used mirror images for this purpose. By measuring learning rates and accuracies, we compare our method with three other Q-learning methods, the normal one, the method using mirror images, and the method using images that move randomly. The results show that the methods using images make learning accuracies worse by a few percent than the normal method, and that our method learns more quickly than the other methods. Also, our method is better in learning accuracies than the method using mirror images.

1. はじめに

近年，人工知能の分野の中で特に機械学習が注目されている．機械学習の多くは最適な結果を導き出すために，訓練データと呼ばれる，入力データと出力データの対を元に学習を行う．一方，機械学習の 1 種である強化学習は訓練データを必要としない学習手法である．

強化学習の手法の性能を評価する方法の 1 つに，追跡問題がある．追跡問題とは，フィールド上でハンターエージェントが獲物エージェントを捕獲するための行動を学習していく問題である．

過去に，学習に用いる状態数を削減することで強化学習を高速化する研究 [1] [2]が行われたが，引き換えに学習精度が悪化する結果となっている．

北尾ら [3]の研究では，強化学習の手法の 1 つである Q 学習の学習精度を維持したまま学習速度を高速化させる手法として，追跡問題においてフィールド上に置かれた鏡に映った獲物エージェントの鏡像を学習に用いる方法を提案している．

本研究ではその手法を発展させ，獲物エージェントの点対称の座標に動く像を用いる手法を使って学習の更なる高速化を目指す．実験では，鏡像を用いた Q 学習と点対称に動く像を用いた Q 学習で追跡問題を行い，学習速度と学習精度を比較した．その結果，点対称に動く像を用いた Q 学習は鏡像を用いた Q 学習と同程度の学習精度に抑えつつ更に学習速度を向上させた．

2. 準備

2.1. Q 学習

Q 学習とは強化学習の手法の 1 つである [4]．Q 学習を行うエージェントは現在の状態sにおいて選択可能な行動aの価値$Q(s,a)$に基づいて次の行動を決定する．一般的に，$Q(s,a)$の値が高い行動は選択される確率が高い．エージェントが行動を行うたびに報酬rに基づいて，$Q(s,a)$を以下の式で更新する．

$$Q(s,a) \leftarrow Q(s,a) + \alpha\{r + \gamma \max_{a' \in A(s')} Q(s',a') - Q(s,a)\}$$

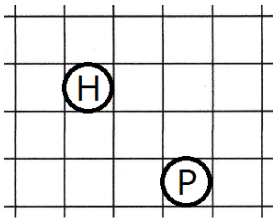
αは学習率と呼ばれ，$0 < \alpha \leq 1$の範囲で設定される．γは割引率と呼ばれ，$0 < \gamma \leq 1$の範囲で，一般的に 1 に近い定数が設定される．$\max_{a' \in A(s')} Q(s',a')$は状態$s$で行動$a$をとった後の状態$s'$において，選択可能な行動の集合$A(s')$の中で最大の価値を持っている行動$a'$の価値である．

2.2. 追跡問題

追跡問題とは，2 次元空間上のフィールドで，ハンターエージェントが獲物エージェントを追跡する問題である(図 1)．ハンターエージェントと獲物エージェントの初期位置はランダムに決定される．全ての価値$Q(s,a)$には初期値として小さな値の乱数が与えられる．エージェントは上に1マス移動，下に1マス移動，右に 1 マス移動，左に 1 マス移動，現在の座標に停滞の 5 つの行動のいずれかを選択する．ハンターエージェントと獲物エージェントは同時に行動し，同一の座標に移動してしまった場合は，どちらのエージェントも行動の選択をやり直す．ハンターエージェントは獲物エージェントとの相対座標を現在の状態sとし，行動を選択する．ハンターエージェントはε-グリーディ法で次の行動を決定する．ε-グリーディ法とは，小さな確率εでランダムに行動を決定し，

Supervisor: Prof. Hiroshi Hosobe

$1-\varepsilon$の確率でその状態で最も Q 値の大きい行動を選択する方法である．



H：ハンターエージェント

P：獲物エージェント

図 1 追跡問題

エージェントの行動後，ハンターエージェントと獲物エージェントが隣接しているかを判定し，行動前のハンターエージェントと獲物エージェントとの相対座標を状態s，ハンターエージェントが選択した行動をaとして，$Q(s,a)$を更新する．その際，エージェント同士が隣接していた場合(図 2)はハンターエージェントが獲物エージェントを捕獲したとして報酬$r=10$，隣接していなかった場合は$r=-1.0$で Q 値を更新する．エージェントの行動と Q 値の更新を終えるまでを 1 ステップとし，ハンターエージェントが獲物エージェントを捕獲するまでを 1 エピソードとする．エピソードを繰り返すことで，ハンターエージェントは獲物エージェントを捕獲するために最適な行動を学習していく．

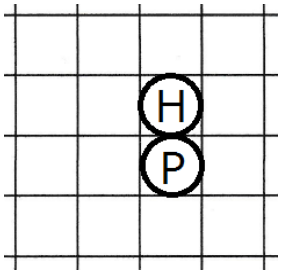
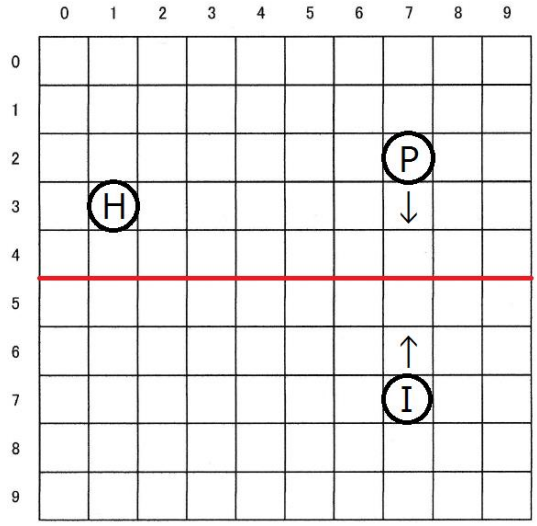


図 2 エージェント同士が隣接した状態

3. 鏡像の利用

本研究は，北尾らの「鏡像を利用した追跡問題の高速学習」[3]を先行研究とする．先行研究では，Q 学習を用いた追跡問題において，ハンターエージェントが現在の状態を観測する際，獲物エージェントとの相対座標だけでなく，獲物エージェントの位置座標を鏡写しに反転させた鏡像との相対座標も用いる手法を提案している．例えば，10×10 マスのフィールド上の中心から横に一直線に鏡を置いた場合，獲物エージェントの座標が$(7,2)$のとき，獲物エージェントの鏡像の座標は$(7,7)$になる(図 3)．この場合，鏡像は左右には獲物エージェントと同じ方向に移動するが，上下には逆の方向に移動する．



H：ハンターエージェント

P：獲物エージェント

I：獲物エージェントの鏡像

図 3 鏡像を用いた Q 学習

ハンターエージェントは移動する際，獲物エージェントとの相対座標を現在の状態として行動を選択する．鏡像がハンターエージェントまたは獲物エージェントと同一の座標に移動した場合は，行動をやり直さず，そのまま学習を続ける．エージェントの行動後，ハンターエージェントと獲物エージェントとの相対座標，ハンターエージェントと鏡像との相対座標の 2 つを現在の状態として 2 回 Q 学習を行う．ハンターエージェントと，獲物エージェントか鏡像のいずれかが隣接していた場合，エピソードを終了とする．この手法を利用することで，Q 学習の学習精度を僅かに悪化させる代わりに，学習速度を向上させることができる．

4. 提案手法

本研究では，フィールドの中心に対して点対称に動く獲物エージェントの像を利用して追跡問題の学習を高速化する手法を提案する．この手法では，エージェントの行動後に，ハンターエージェントと獲物エージェントとの相対座標と，ハンターエージェントと点対称に動く像との相対座標の 2 つを現在の状態として Q 学習を行う．例えば，10×10 マスのフィールド上で，獲物エージェントの座標が$(1,4)$のとき，点対称の像の座標は$(8,5)$になる(図 4)．点対称に動く像は，獲物エージェントとは上下左右反対方向に移動する．

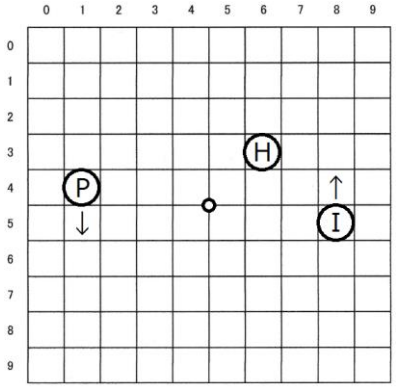


図 4 点対称に動く像を用いた Q 学習

5. 実験

前節の提案手法に基づき，通常の Q 学習，鏡像を用いた Q 学習，点対称に動く像を用いた Q 学習，ランダムに動く像を用いた Q 学習の追跡問題を行い，学習速度と学習精度を計測するプログラムを作成した．使用したプログラミング言語は Java である．本プログラムを用いて以下の実験を行った．

5.1. 方法

複数の手法を利用して，10×10 マスのフィールドで追跡問題の Q 学習を行い，学習速度と学習精度を比較する．学習率$\alpha = 0.1$，割引率$\gamma = 0.9$，確率$\varepsilon = 0.2$とする．

学習結果が収束したとみなす条件を設定し，その条件を満たすまでのエピソード数を学習速度と定義する．それぞれの手法で Q 学習を 20 万エピソード行い，100 エピソードごとの標準偏差を求める．標準偏差が規定値を初めて下回ったエピソードを学習結果が収束したエピソードとみなす．規定値は通常の Q 学習の収束後の標準偏差に近い 6.0 に設定する．

学習精度は，十分に学習が進み，収束条件を満たした Q 値を使ってハンターエージェントに獲物エージェントを追跡させた際の捕獲までのステップ数であると定義する．学習機能を廃した追跡問題の捕獲ステップ数の 100,000,000 エピソード分の平均値を記録する．更に，Q 学習は新たに学習するたびに収束するステップ数が異なるという性質を持つため，これを 10 セット行い，捕獲ステップ数の平均値をその手法の捕獲までのステップ数とする．

実験する Q 学習の手法は，通常の Q 学習(「通常」と呼ぶ)，フィールド上の中心から横に一直線に鏡を置いた Q 学習(「鏡像」と呼ぶ)，点対称に動く像を用いた Q 学習(「点対称」と呼ぶ)である．また，比較対象として，初期位置がランダムで，選択する行動も獲物エージェントに依らずランダムである像を用いた Q 学習(「ランダム」と呼ぶ)も行う．

更に，それぞれの手法の応用形として，像の数が 3 つになる手法も実験する．実験する手法は，図 5 のように，フィールド上の中心から縦と横に一直線に鏡を置いた Q 学習(「鏡像×3」と呼ぶ)，図 6 のように，点対称に動く像に加えて，フィールドの中心を軸にそれを時計回り，反時計回りにそれぞれ90°回転させた像を用いた Q 学習(「点対称×3」と呼ぶ)，図 7 のように，3 つの像が獲物エージェントに依らずランダムに動く Q 学習(「ランダム×3」と呼ぶ)である．

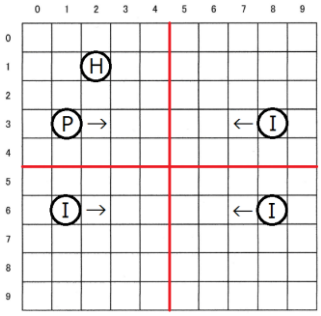


図 5 鏡像×3

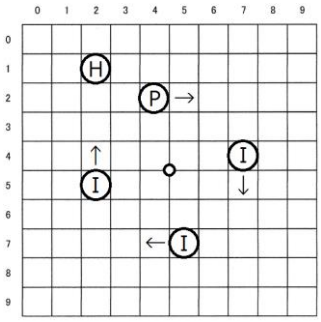


図 6 点対称×3

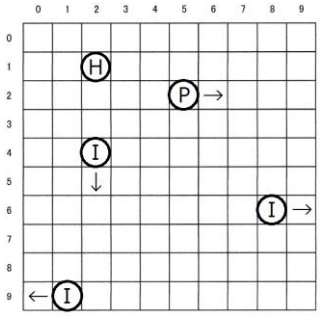


図 7 ランダム×3

5.2. 結果

実験の結果，学習速度は表 1 のようになった．いずれの像を用いた手法も，通常の Q 学習より学習速度が速くなった．また，同じ手法でも，像の数が 1 つのものよりも 3 つのものの方が学習速度は速くなった．点対称に動く像を用いた Q 学習(「点対称」と「点対称×3」)は，鏡像を用いた Q 学習(「鏡像」と「鏡像×3」)や，ランダムに動く像を用いた Q 学習(「ランダム」と「ランダム×3」)よりも高速で学習することができた．

表 1 学習速度

通常	7300 (100.0%)
鏡像	4100 (56.2%)
点対称	1900 (26.0%)
ランダム	4200 (57.5%)
鏡像×3	1500 (20.5%)
点対称×3	1400 (19.2%)
ランダム×3	2200 (30.1%)

学習精度は表 2 のようになった．通常の Q 学習が最も捕獲までのステップ数が少なかった．像の数が 1 つの手法と比較して，像の数が 3 つの手法は学習精度が悪化することがわかった．点対称に動く像を用いた Q 学習は，鏡像を用いた Q 学習や，ランダムに動く像を用いた Q 学習とほぼ同じだが，わずかに良い学習精度となった．

表 2 学習精度

通常	8.4644(100.00%)
鏡像	8.5145 (100.59%)
点対称	8.4723 (100.09%)
ランダム	8.5452 (100.95%)
鏡像×3	8.5473 (100.98%)
点対称×3	8.5335 (100.82%)
ランダム×3	8.5738 (101.29%)

点対称に動く Q 学習は通常の Q 学習，鏡像を用いた Q 学習よりも学習速度が速く，鏡像を用いた Q 学習と同程度の学習精度となった．

6. 議論

点対称に動く像を 1 つ用いた Q 学習が鏡像を 1 つ用いた Q 学習よりも学習速度が速くなったのは，獲物エージェントと像の両方がハンターエージェントから離れる局面が少ないからだと考えられる．図 8 のような局面の場合，獲物エージェントと鏡像がどちらもハンターエージェントから離れてしまうので捕獲までにステップ数がかかり，学習が進みにくいが，図 9 のような局面の場合，獲物エージェントがハンターエージェントから離れると，点対称に動く像は逆にハンターエージェントに近づくので，捕獲までのステップ数が少なくて済むのではないかと考えられる．

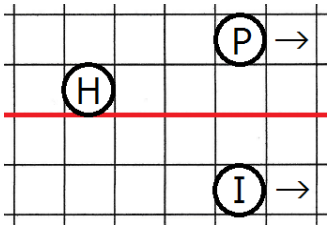


図 8 鏡像を用いた Q 学習で獲物エージェントがハンターエージェントから離れる局面

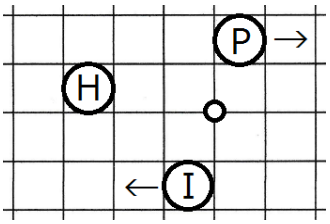


図 9 点対称に動く像を用いた Q 学習で獲物エージェントがハンターエージェントから離れる局面

像を 3 つ用いる手法で，像を 1 つ用いる手法ほど学習速度に差が見られなかったのは，獲物がフィールド全体に分布し，すべての獲物がハンターエージェントから離れる局面が少なかったからだと考えられる．

ランダムに動く像を用いた手法は，獲物の分布に偏りが現れるため，学習速度が向上しにくかったのだと考えられる．

7. おわりに

本研究では，点対称に動く像を用いた追跡問題の Q 学習の手法の提案を行い，通常の Q 学習，鏡像を用いた Q 学習，点対称に動く像を用いた Q 学習，ランダムに動く像を用いた Q 学習の追跡問題における学習速度と学習精度を比較する実験を行った．実験の結果，点対称に動く像を用いた Q 学習は通常の Q 学習よりも，0.1%程度学習精度を悪化させる代わりに，学習速度を約 4 倍に向上させることが分かった．

本研究では，学習率や割引率を固定したが，どのような値が最適であるか議論の余地がある．また，ハンターエージェントが複数存在するマルチエージェントの追跡問題での有用性も検証すべきである．

文　献

[1] 伊藤昭, 金渕満, "知覚情報の粗視化によるマルチエージェント強化学習の高速化―ハンターゲームを例に―," 電子情報通信学会論文誌, vol. J84-D-1, no. 3, pp. 285-293, 2001.

[2] 桑原直哉, 三浦孝夫, "Q 学習による知覚情報の粗視化による追跡動作の学習," 情報処理学会第 73 回全国大会講演論文集, vol. 1, pp. 201-202, 2011.

[3] 北尾健大, 三浦孝夫, "鏡像を利用した追跡問題の高速化," *DEIM Forum*, pp. C7-5:1-5, 2016.

[4] R. S. Barto and A. G. Sutton, 強化学習, 森北出版, 2000.