視覚的顕著性モデルを用いた汎用的機械学習法

野村直也[1,a)]　橋本剛[1,b)]

概要：近年, ゲーム AI 開発は深層強化学習が注目されており, 従来に比べ高い汎用性を持った AI の開発が進んでいる. 代表的なものとして Deep Q Network(DQN) がある. DQN を始めとする深層学習を用いた汎用的なゲーム AI はゲーム画面のみを入力とすることでゲームのルールにとらわれない汎用性を獲得した. 様々なゲームに適用可能であるが, Atari2600 の Ms. Pac-Man のようにあまりスコアの増加が見られないゲームもある. 本稿は人間の視覚における性質をもとに従来の深層強化学習法が苦手としていたゲームへの適応法を提案する. DQN は入力をゲーム画面全体としているが, 人間は画面の中でも注視している部分とそうでない部分があり, 情報の鮮明さが違う. 人間は画面内の重要な箇所を注視して情報を取り入れることで効率的な学習を図っていると考えた. 本稿は Ms. Pac-Man を対象として, 視覚的顕著性モデルを使用してゲーム画面内の重要な情報を抽出することにより学習の効率化を図る. 実験の結果, 学習序盤でより早く高いスコアを取るようになったことが確認できた.

General machine learning using visual saliency models

Nomura Naoya[1,a)]　Hashimoto Tusyoshi[1,b)]

Abstract: In recent years, deep reinforcement learning is gaining attention in game AI development, and development of AI with higher general versatility than conventional ones is progressing. A representative one is Deep Q Network (DQN). A general-purpose game AI using deep learning including DQN gained versatility that is not limited by game rules by inputting only the game screen. Although it is applicable to various games, there are games in which scores are not increased so much like Ms. Pac-Man in Atari 2600. In this paper, we propose a method to improve learning in games which was difficult with conventional deep reinforcement learning method, based on human visual characteristics. Although the input of DQN is the entire game screen, human beings have places they often see in the screen and places they do not, and the sharpness of information is different. We thought that human beings are learning efficiently by paying attention to important places in the screen and extracting information. In this paper, we try to improve learning by extracting important information in game screen using visual saliency models. As a result of experiments on Ms. Pac-Man in Atari 2600, the agent with proposed method got a higher score earlier in the learning process.

1. はじめに

近年, ゲーム AI 開発は深層強化学習が注目されており, 従来に比べ高い汎用性を持った AI の開発が進んでいる. 深層学習を用いることでゲーム画面のような大きな入力を扱うことが容易になり, 人間のような目からの情報を基にした学習プロセスを実現できた. Mnih らは Q 学習と深層学習を組み合わせた Deep Q Network を提案し, 実験では Atari2600 の複数のゲームにおいて, ルールを一切知らない状態から人間よりも高いスコアを獲得するまでの成長を見せた**??**. さらに Hasselt らにより提案された Double Deep Q Network(DDQN) は, DQN の誤差の増大を抑え過度に価値が評価される現象を抑制し, 平均して DQN より高いパフォーマンスを見せた**?**. さらに派生手法が数多く提案されており, 汎用性を獲得しつつもより強力になっている.

しかしゲームによっては人間を超えるほど強くなるものの, 全てのゲームを上手く学習できるわけではない. Mnih

[1] 松江工業高等専門学校
National Institute of Technology, Matsue College
[a)] s1719@matsue-ct.jp
[b)] hashimoto@matsue-ct.jp

らの実験では Atari2600 の中で Ms. Pac-Man や”Montezuma's Revenge”などのゲームでは人間には遠く及ばないスコアしか獲得できなかった?. DQN はゲーム画像を深層学習により解析し, そこにランダムな試行を繰り返すことで各状態に対する最適な行動を学んでいく. このとき画面遷移の幅が小さいほど学習は進みやすくなる. ある状態の最適行動を学んでも画面変化のランダム性が高いと学習された経験が活かされにくくなる. Ms. Pac-Man では目の前の敵に接触し死亡した経験を得ても, その状況が活かされるのは敵が似たような配置にあるときであり, 異なった場所にいると別の状態と識別される可能性が高くなる. このように画面内の変化するオブジェクトやそれらのランダム性に対応させる手法が必要である.

DQN は Atari2600 の複数のゲームで人間を上回る成果を見せた. しかし DQN と違い, 人間は全くプレイできないゲームはなくほぼ全てのゲームである程度のスコアを獲得する. ランダム性が高いオブジェクトが存在するゲームでも, DQN より短い時間でゲームの特徴を掴んでいく. 人間のゲームプレイを考えると, 人間の入力のとり方は DQN と違いがある. DQN ではゲーム画面全ての画像を入力としているが人間は画面全体は見ていない. 人間は目をむけ注視することで先の物体からより多くの情報を得る. ゲーム画面においても同様であり情報を鮮明に取得している箇所とそうでない箇所が存在する. 画面内の重要な箇所の情報を抽出しているため効率的にゲームの特徴を獲得できるのだと考えられる.

画像解析の分野では, 画像内の人間が見る場所を推定する研究が盛んである. 画像の色やエッジなどの特徴量から計算した人間の注視度を顕著性と言い, このモデルから得られる顕著性マップからは画素ごとに人間が注視する確率を得ることができる?. 数多くの手法が存在するが, それらの手法をゲームに適用させる研究は少ない. 本稿では顕著性マップを用いた入力画像の前処理方法を提案することで更に人間に近く汎用的な学習方法を検討する. 深層強化学習のうち DQN と DDQN を用いて Atari2600 の Ms. Pac-Man を対象として実験を行い, 提案手法の優位性を検討する.

2. 深層強化学習

強化学習とは, エージェントが与えられた環境を観測して, そこに行動を加えた結果を元に, 価値が最大化する行動を学んでいく学習プロセスである. この強化学習に深層学習の手法を組み合わせたものが深層強化学習である. ここでは深層強化学習の代表的な手法であり, 本稿で用いている Deep Q Network と Double Deep Q Network について紹介する.

2.1 Deep Q Network

Mnih らにより提案された Deep Q Network(DQN) は強化学習の 1 つである Q 学習における行動価値関数を多層畳み込みニューラルネットワーク用いて近似した手法である?. 深層学習を用いることでゲーム画像のような次元の大きな入力でも上手く学習できるようになった. DQN はゲーム画像だけを入力としており, ゲームのルールを教えられなくても画像の変化と報酬から最適方策を策定することができるため, 従来以上の汎用性を持っていると注目を浴びた.2015 年の Mnih らによる実験では, Atari2600 のゲームの多くで人間レベルのプレイを見せ, いくつかのゲームでは人間を上回るスコアを獲得した?. DQN は様々な方法で改良され, Double Deep Q Network や Dueling Deep Q Network などのより強力な手法も提案されている??.

2.2 Double Deep Q-Network

DQN は, 誤差計算時に目標値が大きすぎると, 前の状態を過大評価してしまうという問題があった. Hasselt らにより提案された Double Deep Q Network(DDQN) は, DQN における行動価値観数 Q を $Q_{\theta^{-1}}$ とそのコピーの Q_θ で分け,s_{t+1} でとるべき行動を Q_θ で決定し, その評価値を $Q_{\theta^{-1}}$ で出力して Q_θ を更新した?. これにより DQN の誤差の増大が抑えられ過度に価値が評価される現象が抑制された. 実験では Atari2600 の多くのゲームで DQN より高いパフォーマンスを見せた.

3. 深層強化学習の難しさ

DQN や DDQN は Atari2600 の多くのゲームで人間よりも高いスコアを獲得した. しかしゲームによっては人間レベルに達しないものもあり, 全てのゲームで上手くいくわけではない. ゲームによりスコアの増加に違いが現れる要因として, ゲーム画面の情報量が挙げられる. DQN や DDQN はゲーム画像を深層学習により解析しランダムに試行することで, その画像の状態に対する最適な行動を学んでいく. このときゲーム画面の変化幅が大きいほど状態数が増えていき, 学習は進みにくくなる. 学習が上手く行く例として, Mnih らの Atari2600 を用いた実験で高いスコアを獲得した Breakout を挙げる. 図 1 は Breakout のゲーム画面上で変化するオブジェクトを示したものである. Breakout では画面上で変化するものはバーとブロックとボールのみである. 画面内のオブジェクトが少ないため, 状態数も少ない. また Breakout はオブジェクト数だけでなくランダム性も少ない. Breakout では図 1 の状況で自分が操作するバー以外に変化するものはボールのみである. このボールにランダム性はなくボールの進む方向に変化するだけであるので, この後に観測される状態はごくわずかに限られる. このように状態変化の幅が小さいゲームでは学習が進みやすく, 短い時間でも高いスコアを獲得しやすい. 対して Mnih らの Atari2600 を用いた実験でスコアが低かったゲームとして Ms. Pac-Man を挙げる. 図 2 は Ms. Pac-Man のゲー

ム画面で変化するオブジェクトを示したものである．Ms. Pac-Man は画面上をプレイヤーが操作するパックマンと敵キャラクターが最大 4 対動く．Breakout よりもオブジェクト数が多く，状態数も多い．また，この敵はある程度ランダム性を持って動くためこの状態から遷移する状態の数は Breakout に比べ遥かに多くなる．オブジェクトが多く，さらにランダム性も加わると，ゲーム内で同じ状況が繰り返されることは難しくなり，学習が進みにくくなる．図 2 では画面下にいるパックマンが水色の敵のすぐ右隣に位置している．このとき次の瞬間に左に進み敵に接触し死んだと仮定すると，次に同じ画面下でパックマンと水色の敵が接触しそうになっても，その他の赤・黄・ピンクの敵の配置が異なれば別の場面と認識される確率が高くなる．ゲームにおいて重要な情報が，その他の重要度の低い情報に引っ張られて学習が進まなくなってしまう．Ms. Pac-Man のようなゲームでは目の前にいる敵に接触し死亡した，または目の前のアイテムを取得すると報酬が増加したなどのゲーム攻略に重要な情報を抽出して記憶することができなければ攻略することは難しい．

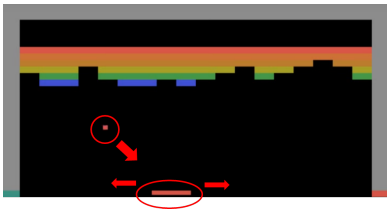

図 1　Breakout の画面変化

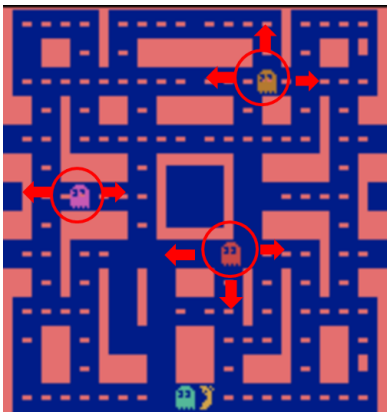

図 2　Ms. Pac-Man の画面変化

4.　関連研究

Yuezhang らにより，オブジェクト認識処理を深層強化学習に適用させる試みがあった?．Atari2600 のいくつかのゲームを対象として，オブジェクトの情報を入力に付加してより高いパフォーマンスを実現させた．画面内のオブジェクト画像を予め用意しテンプレートマッチングにより入力画像内のオブジェクトを抽出する．そしてイメージ内で k 個のオブジェクトが検出されたとすると，k 個の追加のチャネルを元のイメージの RGB チャネルに追加する．各チャネルは単一のオブジェクトを表し，各チャネルでは検出されたオブジェクトに属するピクセルに対し，対応する位置に 1 を割り当て，そうでない場合は 0 を割り当てる．これによりオブジェクトの位置と属性を入力に付加する．このオブジェクト情報を加味した深層強化学習は，オブジェクト情報を付加しないものよりも優れたパフォーマンスを示した．5 つの Atari2600 ゲームで実験が行われ，各ゲームで 1 %から 20 %スコアが上昇した．Ms. Pac-Man では対象のゲームの中で最も高い 20 %のスコアの上昇が見られた．

オブジェクトを抽出することでオブジェクトの場所や属性といったゲーム攻略に重要な情報を抽出でき，これにより効率的に学習が進んだのだと言える．しかし，Yuezhang らはオブジェクト情報の抽出は各ゲームのオブジェクト画像を用いたテンプレートマッチングに拠るものであった．これは各ゲームに個別のタスクとなっており，汎用性という観点ではあまり優れていない．

5.　人間のゲームプレイ

では人間は Ms. Pac-Man のようなゲームをどのようにプレイしているのだろうか．人間のゲームプレイを考えると，人間と DQN や DDQN には入力のとり方に違いがある．DQN や DDQN はゲーム画面全体を入力としているのに対し，人間は画面全体は見ていない．Ms. Pac-Man であれば，自機キャラクターや近い場所にいる敵，近くのアイテムなどをよくみてプレイするであろう．遠くの敵はぼんやりと見ており，自機の近くを注視することが多いと考えられる．このように人間にはゲーム画面内でも状況によりよく見る箇所とそうでない箇所が存在し，重要でないと判断した部分の情報を削り，重要な箇所の情報を鮮明に取り入れることで情報量が多いゲームでも効率的に学習していると考えられる．

6.　視覚的顕著性

人間は，目に写った映像から重要だと思われる箇所を瞬時に判断してその部分を注視している．この機能をコンピュータ上で実現するための研究が盛んに行われており，様々な計算モデルが存在する．様々な画像特徴量をもとに計算され推定された「注視される確率を表す画像」を顕著性マップと呼ぶ．Itti らが提案した計算モデルが最も広く知られており，初期視覚系の生理学的知見に基づき，色や傾きといった基本的な視覚属性の顕著性，すなわち目立ちやすさの分布を計算し，これに基づいて顕著度の計算を行っている?．この手法は様々な形で改良されており，グラフベースのアプローチに拠るモデルが提案されている．Hou らによるモデルは用いる特徴は Itti らと同様であるが，対象物体で

はなく背景の特徴に着目して, 周波数スペクトルを用いている?. 近年では動画像における視覚的顕著性や, オブジェクトを識別することを目的とした物体認識にも応用されている.

このような手法はゲームに適用された事例は少ない. 本稿では視覚的顕著性を求める計算モデルのうち, スタンダードとされる Itti らの計算モデル, 及び Itti らを基にした Hou らのモデルを DQN, DDQN に組み込むことで, 学習の効率化を図る.

7. 提案手法

本稿では顕著性マップを用いた入力画像の前処理方法を提案する. まずゲーム画面画像を取得後, 顕著性計算モデルを用いて顕著性マップを作成する. 顕著性マップは画素ごとに注視されやすさが数値として表現されており, 一定値以上もしくは最大値を注視部分とする. 注視されると推定された箇所以外にはフィルタをかけぼかし効果を入れ情報を落とす. フィルタは注視箇所から離れるごとに強くしていき, 遠くの情報ほど多く減らしていく. 注視箇所が自機周辺の場合, 近くの敵の情報は鮮明に残るが, 遠くの敵やアイテムの情報はフィルタによりぼやけたものとなる. この処理を施した画像を従来の深層強化学習に入力として与える. この処理を加えることで, 画像中の変化が抑えられ学習がより進むと考える. この処理を加えた DQN 及び DDQN を本稿では Saliency DQN(S-DQN), Saliency DDQN(S-DDQN) とする.

8. 事前実験

顕著性計算モデルを用いて正しく注視箇所を推定できるか実験を行った. Ms. Pac-Man において自機が最も重要であり注視されるべき箇所であると考え, どの計算モデルが最も自機を注視箇所と推定できるか調べた. 計算モデルは Itti らのモデルと Hou らのモデルを使用した. Ms. Pac-Man のゲーム画像をランダムに 100 枚用意し, モデルよりそれぞれの顕著性マップを作成した. そしてマップ上の注視確率が 0.9 以上の画素を注視箇所としてその画素を中心とする半径 20 の円を描画した. その後目視により 100 枚の画像の内, 自機キャラクターを含む箇所に円を描画している画像を調べた. 図 3, 図 4 は実験結果の一部である. 図 3 では画面中央のパックマンが, 赤い円で囲まれている. 赤い円の中心が注視画素であり, 赤い円の付近を見る確率が高いことを表している. 図 4 は画面内の離れた複数の箇所が注視点と推定されたゲーム画像である. 状況により自機だけでなく敵も赤い円で囲まれ, あまり人間が見るとは考えられない画面下の部分も注視点と推定されている. 実験の結果, Hou らのモデルでは全ての画像で自機を含んだ箇所を注視箇所と推定できていた. 画像により自機が抽出されないこともあったが 100 枚中 96 枚で円の内に自機が含まれていた. Itti らのモデルは自機を注視箇所としたものは半分ほどであった.

また Hou らのモデルにおいて, 顕著性マップの最大値を注視箇所として実験を行った. 図 5 は自機が円に含まれているが, 図 6 は離れた敵に円が描かれている. 最大値だけを取ると, 確率 0.9 以上を注視箇所とした時に比べ自機が含まれないことが多かった. 結果として 100 枚中 70 枚の画像で自機を抽出していた. 以上の結果より, 顕著性計算モデルは Ms. Pac-Man では Hou らの計算モデルが優れていることが確認できた. 最大値を注視点とすると, 自機が含まれない可能性が高くなったが, 70 %の確率で自機を注視点と推定できていた. ただし計算モデルを適用し, 顕著性マップから注視点を推定するまでの実行時間は, 最大値のみの抽出の場合, 注視点を複数設けた場合に比べ 3 倍以上高速であった.

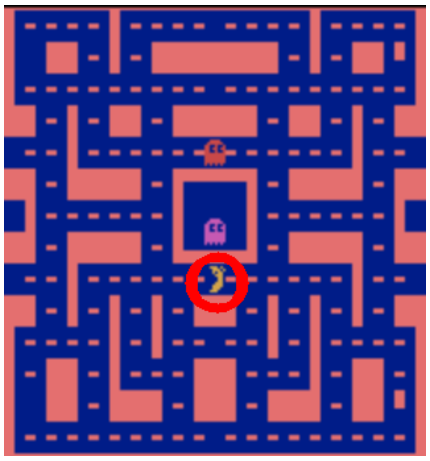

図 3 Ms. Pac-Man の画像に視覚的顕著性モデルを適用し, 注視確率 0.9 以上の画素を中心に円を描画した画像 (自機付近のみが抽出された場合)

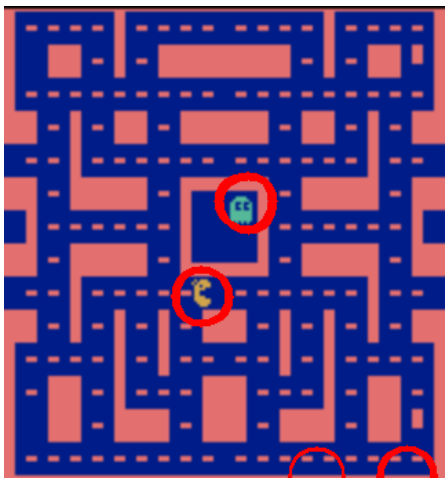

図 4 Ms. Pac-Man の画像に視覚的顕著性モデルを適用し, 注視確率 0.9 以上の画素を中心に円を描画した画像 (離れた複数箇所が抽出された場合)

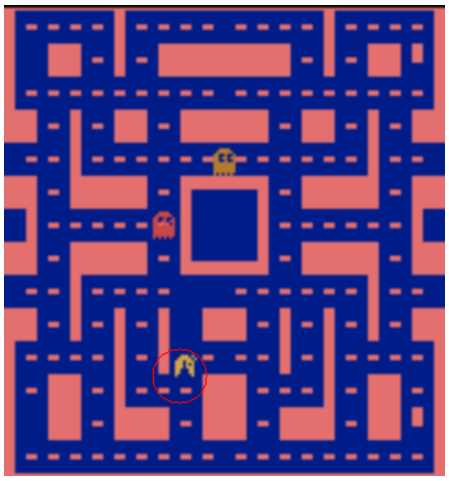

図 5 Ms. Pac-Man の画像に視覚的顕著性モデルを適用し, 注視確率が最大値の画素を中心に円を描画した画像 (自機のみが注視箇所と推定されている場合)

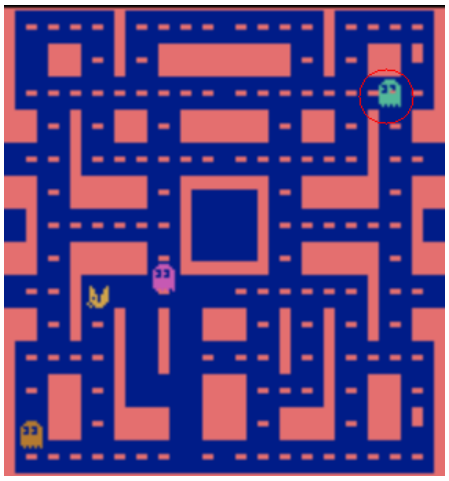

図 6 Ms. Pac-Man の画像に視覚的顕著性モデルを適用し, 注視確率が最大値の画素を中心に円を描画した画像 (自機以外が注視箇所と推定されている場合)

9. 実験

表 1 に実験環境を示す. また, 表 2 に用いた内部手法及びパラメータの値を示す. Chainer 及び Chainer の強化学習用ライブラリである ChainerRL, そして強化学習用プラットフォームである OpenAIgym を使用して実験を行った [1]. 実験対象として Ms. Pac-Man を選択した. Ms. Pac-Man は Mnih らの実験において Atari2600 のゲームの中でスコアが低く, 十分な学習ができていなかったためである?. 学習手法は DQN, DDQN, S-DQN, S-DDQN を用いた. ネットワーク構造は 2015 年の Mnih らの実験と全て同じものを使用し, 表に示すように Mnih らの実験と同様のものを使用した?. 学習プログラム内部で使用したパラメータの値についても同様である.

顕著性計算モデルは第 8 章で述べたように Hou らのモデルを使用し, OpenCV の saliency モジュールを使用した. 注視箇所は顕著性マップの最大値とした. 注視箇所を複数箇所とすると, 最大値の一点のみとしたものより学習時間が約 3 倍長くなったため, 時間の都合上本稿ではより高速な最大値のみを注視点と推定する方法で実験を行った. 顕著性モデルより得られた注視画素を中心とした縦横 30 ピクセルをフィルタをかけずに入力とした. その範囲から縦横 30 ピクセルずつ拡大し 5 × 5 のフィルタを用いてガウシアンフィルタを適用した. さらに縦横 30 ピクセルごとにフィルタをかける範囲を拡大していく. フィルタ範囲を拡大するごとにフィルタサイズを 10 × 10, 15 × 15 と増やし, より情報をぼかしていった. 図 7 のゲーム画面画像に処理を加えたものが図 8 である. 図 8 のような画像を状態として与え, 直近 4 フレームの画像を畳み込みニューラルネットワークに入力した. ゲーム開始から死ぬまでを 1 ステップとし, 用いた 4 手法でそれぞれ 15000 ステップ学習させた. また,50 ステップごとに 10 回のテストプレイをさせその平均スコアを記録した. また, 学習後のモデルを使用して Ms. Pac-Man を 50 回プレイさせそのスコアを記録した.

表 1 実験環境

OS	Ubuntu16.04LTS
GPU	G-Force GTX 1070
言語	Python2.7
ライブラリ	Chainer4.3.0
	ChainerRL0.3.0
	OpenCV3.3.0

表 2 パラメータおよび使用手法

割引率	0.99
学習係数	0.1
行動決定	ε-Greedy 法
活性化関数	ReLU
optimizer	RMSprop

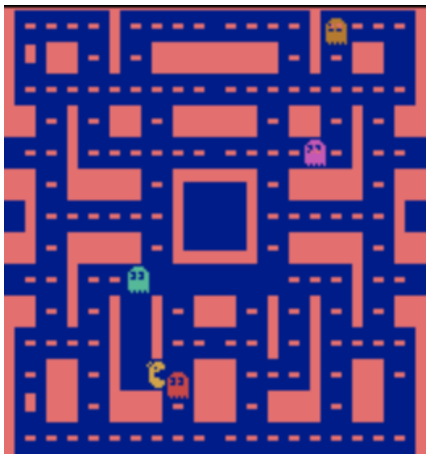

図 7 Ms. Pac-Man のゲーム画像

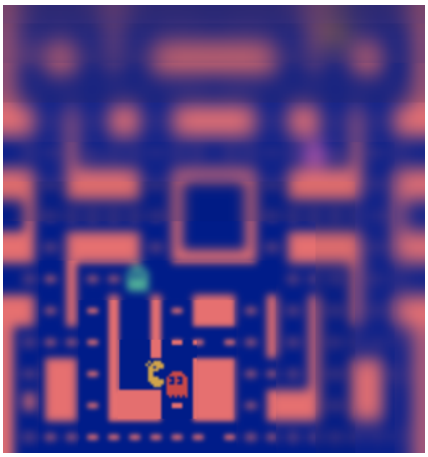

図 8 図 7 に提案する前処理を施した画像

10. 結果

図 9 学習中の報酬推移を表している. 横軸は学習ステップ数で, 50 ステップごとに 10 回テストプレイを行った. 縦軸は 10 回のテストプレイの平均報酬である. グラフは区間 10 で移動平均を取ったものである. 15000 ステップの学習には, DQN と DDQN は約 2 日間, S-DQN と S-DDQN は約 5 日間の時間を要した. また, 表 3 は学習済モデルに Ms. Pac-Man を 50 回プレイさせた時の平均スコアである. 図 9 では学習序盤は S-DQN, S-DDQN が従来の DQN, DDQN に比べ早い段階でスコアが増加した. その後 DQN 及び S-DQN では 10000 ステップのあたりでスコア上昇が停滞した. DDQN 及び S-DDQN はどちらも 5000 ステップの付近から停滞している. 表 3 では DQN と DDQN では 100 点近くの差があるが, DQN と S-DQN, そして DDQN と S-DDQN はどちらもスコアの大きな違いは見られなかった.

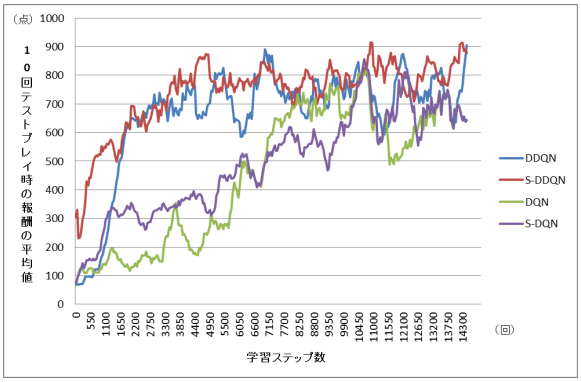

図 9 15000 ステップ学習時の 50 ステップごとのテストプレイにおける平均報酬 (区間 10 で移動平均をとったグラフ)

表 3 学習後モデルに Ms. Pac-Man を 50 回プレイさせたときの平均スコア

DQN	S-DQN	DDQN	S-DDQN
659	627	754	789

11. 考察

図 9 より, DQN, DDQN に比べ, S-DQN, S-DDQN はどちらも早い段階から報酬が増加している. 従来の DQN と DDQN は序盤に少し停滞し,1000 ステップのあたりで徐々に報酬が伸びているのに対し, 提案手法を組み込んだものはどちらも学習開始からすぐに報酬が増加している. これは提案手法により入力であるゲーム画像の変化が抑えられ学習が効率的に進んだためではないかと考える. DQN や DDQN はゲーム画面だけを入力として学習するため, 画面の情報を減らすと学習が進まなくなるおそれがある. しかし S-DQN と S-DDQN は画面の情報を減らしても従来と同程度の強さになった. これは視覚的顕著性モデルを用いて重要な情報を正しく抽出することができたためだと考える.

S-DDQN は学習中盤以降の報酬は増加が止まり, 従来の手法に追いつかれている. Yuezhang らの Ms. Pac-Man を対象とした実験でも, 用いた内部ネットワークやパラメータが違うため本稿とスコアに差があるが, どの手法も 200 万フレームあたりで学習が停滞していた. このように DQN のような学習手法では Ms. Pac-Man の学習にはスコアの上昇がある程度で停滞する傾向があると思われる.

本実験では学習時間の都合上注視点を一点のみとしたため, 自機以外に注視点が選ばれることがあった. 事前実験では注視確率 0.9 以上の点を抽出したところほぼ全てで自機を含んだ箇所を注視点と推定することができていたため, そちらの条件で実験を行うことで, 本実験以上の報酬増加が見られるのではないかと考える.

12. まとめ

本稿では, 視覚的顕著性計算モデルを用いてゲーム画面内の自機や敵などの重要な情報を抽出することで, DQN や DDQN の効率的な学習を促すことができると考え実験を行った. 具体的には Hou らの視覚的顕著性モデルを用いてゲーム画像の注視箇所を推定し, その箇所以外の情報をフィルタによりぼかしてエージェントへの入力とした. 結果として, 従来の手法に比べ学習序盤により早くスコアが上昇した. 本手法により, DQN と DDQN のより早い学習を促進することができたと思われる. 本稿では顕著性マップの最大値のみを注視点としていたため, 注視点を一定値以上として複数設けて重要な情報を全て抽出することで, さらなる報酬が増加すると考える.

13. 今後の展望

注視箇所を複数個として学習させる. また本稿と同じ条件で複数回実験を行い, スコアの推移を確認する必要がある. そして Ms. Pac-Man 以外にも, Mnih らの実験においてスコアが増加が見られなかったゲームを対象として実験を行い学習を促進できるか検証したい.

参考文献

[1] Greg Brockman et al. Openai gym, 2016.

[2] L. Itti et al. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 11, pp. 1254–1259, 1998.

[3] V. Mnih et al. Playing atari with deep reinforcement learning. *NIPS 2013 Deep Learning Workshop*, pp. 2495–2503, 2013.

[4] V. Mnih et al. Human-level control through deep reinforcement learning. *Nature*, Vol. 518, No. 7540, pp. 529–533, 2015.

[5] X. Hou et al. A spectral residual approach. *CVPR*, 2007.

[6] Yuezhang Li et al. Object-sensitive deep reinforcement learning, 2018.

[7] Ziyu Wang et al. Dueling network architectures for deep reinforcement learning, 2015.

[8] H. van Hasselt et al. Deep reinforcement learning with double q-learning. *CoRR*, 2015.