F-042

可視化による Deep Q Network の行動価値根拠の分析
Analysis of the Action Value of Deep Q Network by visualization

長嶺 一輝† 遠藤 聡志† 山田 孝治† 當間 愛晃† 赤嶺 有平†
Kazuki Nagamine Satoshi Endo Koji Yamada Naruaki Toma Yuhei Akamine

1. はじめに

人間が視覚情報を元にタスクを解くとき，そのタスクに出現するオブジェクトなどの視覚的特徴に注視する[1, 2]．映像のみを入力に，Atari 2600 [3] という 2D ゲーム環境において人間並みのパフォーマンスを発揮した，Deep Q Network [4] という深層強化学習アルゴリズムがある．このアルゴリズムにおいても，人間と同じようにゲーム内のオブジェクトに注目するといった現象が起きていると推測できる．本研究では，DQN が行動価値を計算する際に用いる CNN に Grad-CAM [5] という手法を用いる．これによって，価値の根拠となる入力領域を可視化し，DQN が学習過程で捉えた特徴を 分析する．

2. 関連研究

2.1 Deep Q Network (DQN)

Deep Q Network (DQN) [4] は，Atari 2600 [3] のいくつかの環境において，人間と同等以上の性能を記録した深層強化学習手法である．深層強化学習とは，強化学習に深層学習を組み合わせる手法で，DQN はその一つである．DQN は古典的な強化学習手法の Q 学習を深層学習で拡張したもので，行動価値関数をディープニューラル ネットワークで関数近似している．特に，DQN ではネットワークに畳み込みニューラルネット (CNN) を用いている．これによって，画像入力に対して，問題を解くために必要な特徴を自動的に抽出している．この特徴から行動価値を計算し，ランダムな動きも取りながら学習を進める，

2.2 Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) [5] は，学習済み CNN の入力画像に対する特徴部位を可視化する手法である．Grad-CAM は Class Activation Mapping (CAM) [6] を拡張したもので，CAM が CNN の最終層に Global Average Pooling を必要とするのに対して，Grad-CAM はその必要なく使用することができる．あるクラス c に対して Grad-CAM を用いてヒートマップ $L^c_{Grad-CAM}$ を求める式を以下に示す．

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

$$L^c_{Grad-CAM} = ReLU\left(\sum_k \alpha_k^c A^k\right) \tag{2}$$

ここで，(1) 式の A は CNN のある層の出力である特徴マップ，Z はそのサイズ，k はその番号，y は出力層の出力ベクトルである．このように，クラス毎に出力への影響が大きい入力領域を，微分係数の平均化を用いて求めている．

3. 提案手法

DQN の行動価値の根拠となる入力領域を可視化するために，以下のプロセスを提案する．

(1) ゲーム環境を解くために DQN を学習させる
(2) 学習前後の CNN の重みを保存する
(3) (2) で保存した重みを使って CNN を再構築する
(4) (3) を用いて環境の状態を入力として出力と特徴マップを得る
(5) (4) で得た出力と特徴マップを用いて Grad-CAM を計算し，ヒートマップを得る

DQN の持つ CNN は，行動価値を出力するために最終層の出力に線形回帰を用いている．これでは Grad-CAM の出力を正しく得られないため，(3) のプロセスで CNN を再構築したあと，出力にソフトマックス関数を用いるように変更した．また，各行動についてヒートマップを得るために全ての行動について Grad-CAM を行う．図 1 に提案手法の各プロセスを示す．

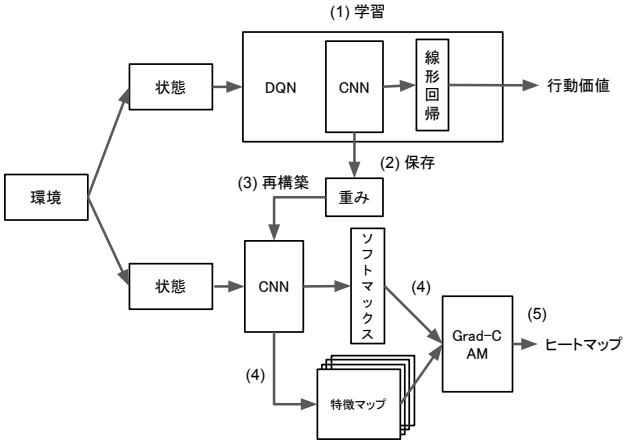

図 1　提案手法の流れ

4. 実験

4.1 実験概要

計算機実験では提案手法に則り，ゲーム環境で DQN を学習しながら CNN の重みを保存する。学習終了後に CNN を再構築して Grad-CAM を使ったヒートマップの出力を行う．学習時間を短縮するために，環境には OpenAI Gym [7] 上に Pygame [8] で作成した Atari の Breakout の簡略化版を構築した．図 2 にその環境から得られる状態画像のサンプルを示す．画像は 1 から 0 に正規化し，サイズは 60×60×3 である．エージェントはこの環境に対して，パドルの左右の加速及び無行動の 3 つの行動を選択することが可能である．また，モデルの学習ステップ数やハイパーパラメータなどを表 1 に示す．

† 琉球大学 University of The Ryukyu,

図 2　Pygame Breakout

表 1　学習に用いた設定

パラメータ名	値
学習ステップ数	400000
Conv1	8×8×30, (4, 4), ReLU
Conv2	4×4×40, (3, 3), ReLU
Conv3	3×3×60, (1, 1), ReLU
Pooling	Global Average Pooling
FC	512
学習率	0.00005
最適化手法	RMSProp

表 1 の Conv は畳み込み層の設定で，数字が小さい方が入力層に近く，値は左からカーネルの高さ・幅・数，ストライドの幅，活性化関数である．畳み込み層の後にプーリングを行い，その後全結合層を接続している．

4.2 実験結果

図 3 に DQN の学習開始から終了までの獲得報酬の推移を示す．報酬は 2000 ステップ毎の平均報酬を記録している．縦軸が平均報酬の値，横軸がステップ数である．学習前後で保存した重みを使って，Grad-CAM から得たヒートマップと入力画像を重ね合わせたものを図 4 に示す．図 4a は学習前の重みを使用したもので，図 4b は学習後の重みを使用したものである．また，4a，4b のそれぞれの画像は左から順にパドルの左加速，右加速，無行動の各行動に対応する Grad-CAM の出力である．

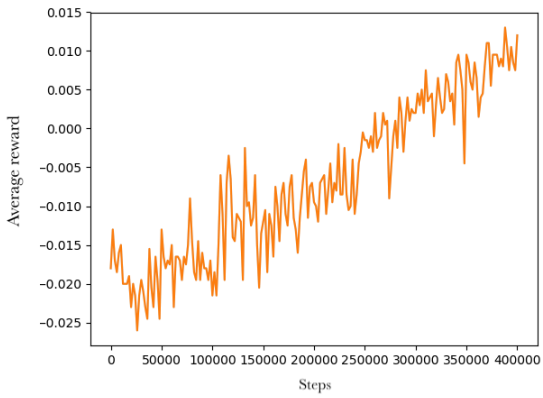

図 3　DQN の平均獲得報酬の推移

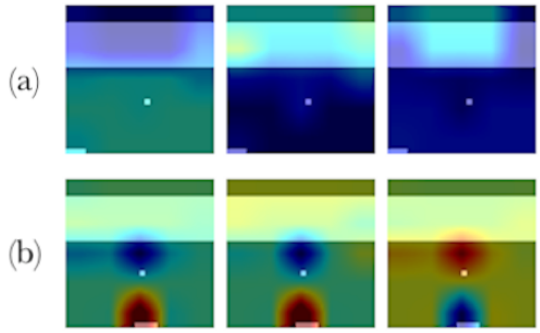

図 4　Grad-CAM 適用後の入力画像

4.3 結果の考察

図 3 から DQN の学習が進んでいることがわかる．図 4a では，Grad-CAM のヒートマップが一様に青く，ボールやパドルといったオブジェクトに注目できていない．図 4b では，オブジェクトの周りが顕著に強調されている．また，学習後の重みを用いた図 4b では，左右加速ではパドルが赤く強調されて出力に対して正の貢献をしており，ボールが青く強調されて負の貢献をしているのに対して，無行動では反対のことが起きているのが見て取れる．

5. まとめ

本研究では，DQN の行動価値の根拠となる入力領域を，DQN が持つ CNN に Grad-CAM を用いることで可視化し，結果を分析した．その結果，Breakout の中で重要なオブジェクトであるボールやパドルを根拠としてハイライトしていることを視覚的に分析した．このような注目は，私たち人間も同様に行なっていると考えられる．今後の課題として，提案した可視化手法を用いて，オブジェクト間の位置関係といった特徴が取れているか調査することや，他の深層強化学習手法や環境に対して同手法を適用することがあげられる。

参考文献

[1] Connor, Charles E., Howard E. Egeth, and Steven Yantis, "Visual attention: bottom-up versus top-down.", Current biology 14.19 (2004): R850-R852.
[2] Rensink, Ronald A, "The dynamic representation of scenes.", Visual cognition 7.1-3 (2000): 17-42.
[3] Bellemare, Marc G., Joel Veness, and Michael Bowling, "Investigating Contingency Awareness Using Atari 2600 Games.", AAAI. 2012.
[4] Mnih, Volodymyr, et al, "Human-level control through deep reinforcement learning.", Nature 518.7540 (2015): 529.
[5] Selvaraju, Ramprasaath R., et al, "Grad-cam: Visual explanations from deep networks via gradient-based localization.", IEEE International Conference on Computer Vision (ICCV). 2017.
[6] Zhou, Bolei, et al, "Learning deep features for discriminative localization.", Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on. IEEE, 2016.
[7] Brockman, Greg, et al, "Openai gym.", arXiv preprint arXiv:1606.01540 (2016).
[8] Shinners, Pete, "Pygame." (2011).