The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

深層強化学習エージェントの行動別顕著性マップの生成に関する考察

Consideration on Generation of Saliency Maps in Each Action of Deep Reinforcement Learning Agent

長嶺一輝 *1 Kazuki Nagamine　遠藤聡志 *2 Satoshi Endo　山田孝治 *2 Koji Yamada　當間愛晃 *2 Naruaki Toma　赤嶺有平 *2 Yuhei Akamine

*1琉球大学理工学研究科情報工学専攻
Information Engineering Course, Graduate School of Engineering and Science, University of the Ryukyus

*2琉球大学工学部工学科知能情報コース
Faculty of Engineering, School of Engineering Computer Science and Intelligent Systems, University of the Ryukyus

In recent years, deep reinforcement learning agents have surprisingly developed and achieved great results. the methods of analyzing the behaviour of agents by visualizing neural networks have been proposed. However, the methods to obtain saliency maps for each action has not been much researched. In this paper, we propose the method of generating saliency maps for each action of the agents in order to obtain deeper insight when analyzing a neural network in a deep reinforcement learning agent by visualization. We applied the proposed method to the agent which learned Atari 2600 Pong. As a result of the experiment, we obtained saliency maps which visualizes the influence of environment on each action of the agents.

1. はじめに

近年，深層学習の発展に伴い，強化学習に深層学習を取り入れた深層強化学習も目覚ましい成長を見せている．一方で，深層学習ではブラックボックス的な性質があるため，その解消が課題となっており，深層強化学習にも同様の問題がある．例えば，エージェントの行動のみを視認して，根拠となった画像特徴を推測するのは困難である．これにアプローチする手法としてニューラルネットの入出力を用いて判断根拠を可視化する手法が提案されている [Selvaraju 17]．このような手法は深層強化学習においてはエージェントの行動根拠の視覚化に用いられている．顕著性マップの生成は可視化手法の一つであり，エージェントが注視しているオブジェクトや学習における戦略の変化等の分析に活用され始めている．しかし，エージェントが出力する行動価値セットに対して一つの顕著性マップを得る手法は提案されているが，出力を行動ごとに分けて可視化する手法は十分に研究されていない．行動別に可視化できれば，分析においてより深い洞察が期待できる．そこで，本研究では既存の可視化手法を拡張して，深層強化学習エージェントの行動ごとに顕著性マップを得る手法を提案する．また，提案した手法を Atari 2600 の Pong を学習した深層強化学習エージェントに適用した結果を示し，得られた行動ごとの顕著性マップについて考察する．

2. 先行研究

Greydanus らは深層強化学習エージェントの行動等を分析するために，顕著性マップを用いた可視化手法を提案している [Greydanus 17]．この提案手法と深層強化学習アルゴリズムである Asynchronous Advantage Actor-Critic (A3C) 及び Atari 2600 ゲーム環境を用いて，エージェントが注視している部位を可視化し，学習過程における戦略の変化等について分析している．また，ゲームや機械学習に精通していない非エキスパートでも，可視化結果を見ることでエージェントの行動の解釈が容易になることを示している．Greydanus らは提案手法を顕著性マップを生成する摂動ベースな方法と呼び，次のようにマップを求めている．はじめに，エージェントを十分に学習させた後，学習を停止した状態で環境からの観測やエージェントが出力した行動価値等を数ステップ分保存する．次に，保存した観測状態と次の (1) 式を用いてマスク画像を作成する

$$\Phi(I_t, i, j) = I_t \odot (1 - M(i,j)) + A(I_t, \sigma_A) \odot M(i,j) \quad (1)$$

ここで，Φ はマスク画像，I_t はステップ t における観測状態，i, j はその二次元座標，$\odot$ はアダマール積，M はマスク，A はガウスフィルターでぼかした観測状態を表す．マスク画像は，観測状態の二次元座標 i, j をマスク M の範囲でぼかしたもので，エージェントが行動根拠とする画像特徴量を認識不可にする効果を持つ．これと (2) 式の顕著性メトリックを用いて，画像部位 i, j における顕著性スコアを計算する．顕著性スコアとは，その部位がエージェントの行動価値に及ぼす影響を数値化したものである．

$$S(t,i,j) = \frac{1}{2}\|\pi_u(I_{1:t}) - \pi_u(I'_{1:t})\|^2$$
$$\text{where} \quad I'_{1:k} = \begin{cases} \Phi(I_k, i, j) & \text{if } k = t \\ I_k & \text{otherwise} \end{cases} \quad (2)$$

ここで，S は座標 i, j におけるスコア，π_u は行動価値ベクトル，$I_{1:t}$ は観測状態の系列を表す．このスコアを縦横数ピクセル間隔で求め，顕著性マップを生成する．求めたマップはリサイズし，前処理なしの観測状態の RGB チャンネルのいずれかと加算することで可視化する．上記の (1)(2) 式は A3C のアクターに対する顕著性スコアの計算式で，クリティックについても状態価値を用いて同様に計算することができる．青がアクター，赤がクリティックの顕著性マップとして，Pong 環境下で当手法を用いて可視化した結果を図 1 に示す．図 1 より，Pong をプレイする上で重要と考えられるブロックやパドルといったオブジェクトをエージェントが注視していることが直感的にわかる．

当手法がアクターの行動根拠とするオブジェクトをハイライトしている一方で，どのオブジェクトがいずれの行動価値に貢

連絡先: 長嶺一輝，琉球大学理工学研究科，〒 903-0213 沖縄県中頭郡西原町千原 1，k188583@ie.u-ryukyu.ac.jp

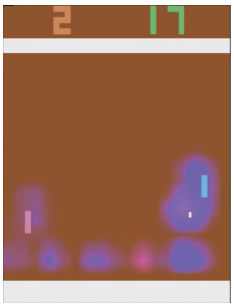

図 1: Greydanus らの提案手法による可視化画像

献しているかといった観察を行うことは難しい．そこで，本研究では (2) 式の顕著性メトリックを変更して，行動別に顕著性マップを得る手法を提案する．

3. 提案手法

本研究では，各行動ごとの顕著性マップを得るために，先行研究の (2) 式の $S(t,i,j)$ を次の (3) 式のように変更した．

$$S(a,t,i,j) = \pi_a(I_{1:t}) - \pi_a(I'_{1:t}) \tag{3}$$

ここで，a は行動価値ベクトルのインデックス，π_a は行動価値ベクトル内の a に対応する一つの行動価値を表す．この変更により，行動別の可視化だけでなく，エージェントが注視している部位が行動価値に対して正負どちらの貢献をしているか可視化することも狙った．観測状態に対する行動価値と，マスク後の行動価値の変化及びマスク部位の行動価値に対する貢献の関係を表 1 に示す．

表 1: 行動価値とその変化とマスク部位の貢献の関係

行動価値	行動価値の変化	部位の貢献
正	増加	負
正	減少	正
負	増加	負
負	減少	正

これは，例えば，観測状態に対する行動価値が正で，マスク後の行動価値が増加した場合，その部位が行動価値に対して負の影響を与えると解釈できることを表す．この表 1 による解釈と (3) 式を用いることでエージェントの行動別顕著性マップを生成する手法を提案する．

4. 実験

本実験では提案手法により各行動ごとの顕著性マップを得ることを目的とする．また，得られた可視化結果と図 1 とを比較する．

実験設定として，本稿では学習環境に Pong を用いる．Pong には Up，Down を含む六つの行動がある．また，結果を解釈し易くするためにオブジェクトの軌跡等の表示を追加した．

実験結果を示す．図 2 は，実験結果の動画から抜き出したフレームで，対戦相手が打ったボールをエージェントがパドルを操作して打ち返そうとしている場面である．図中の茶色と白色の部分が Pong の画像に顕著性マップを合成したもので，赤の部位が行動価値に正の貢献を，青の部位が負の貢献をしてる．

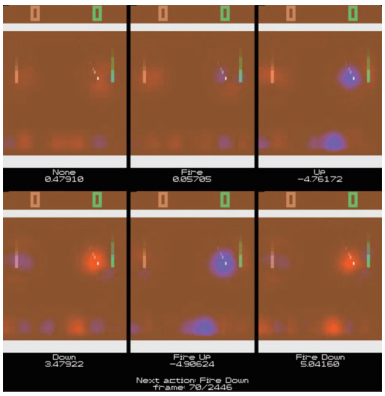

図 2: 本研究の提案手法による可視化画像

それぞれの画像の下には行動ラベルと行動価値を，図中下部にはエージェントが取る行動とステップ数を表示している．各画像の行動価値を見ると，Fire Down が最も高く，次に Down が高いことがわかる．反対に，Fire Up が最も低く，Up が次に低いことがわかる．None と Fire はその中間程度の価値である．各画像では，None と Fire は特徴的な注視は生じておらず，残りの四つに顕著な特徴を持った注視が生じている．一方で，この四つは注視部位がおおよそ同じで，Fire と Down でオブジェクトの貢献が対照的になっていることがわかる．

考察を述べる．ボールの軌跡は，ボールがパドルより下に向かっており，パドルもそれに追従するように下に向かっている．ボールはパドルから距離があり，パドルの x 座標に到達するまで時間がかかることから，パドルを下に操作する Donw，Fire Down の行動価値が高いと考えられる．また，その際の顕著性マップを見ると，ボールが正の貢献をしていることがわかる．反対に，Up，Fire Up においてはボールが負の貢献をしている．これは，各オブジェクトがこの位置関係の時は，ボールが与える貢献が行動価値に大きく影響しているためだと考えられる．オブジェクトと各行動の関係についての洞察を得られることが，提案手法の先行研究に対するアドバンテージと考える．

5. まとめ

本研究では，深層強化学習エージェントの行動根拠を視覚的に分析するために，摂動ベースの顕著性マップ生成を行う既存手法を基に，行動別の可視化を行う手法を提案した．実験では，A3C に提案手法を適用し，Pong 環境下でのアクターの行動別の顕著性マップを生成した．その結果，行動別に特徴的な注視部位が生じていることを確認した．また，結果からオブジェクトが行動に与える影響といった，先行研究の手法では得ることが難しい洞察を得られた．

参考文献

[Selvaraju 17] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, in Proc. of ICCV-17, pp. 618626 (2017)

[Greydanus 17] Greydanus, S., Koul, A., Dodge, J., and Fern, A. Visualizing and Understanding Atari Agents. arXiv preprint arXiv:1711.00138 (2017)