A3C における Attention 機構を用いた視覚的説明

Visual explanation using Attention mechanism in A3C

板谷 英典 *1　平川 翼 *1　山下 隆義 *1　藤吉 弘亘 *1　杉浦 孔明 *2
Hidenori Itaya　Tsubasa Hirakawa　Takayoshi Yamashita　Hironobu Fujiyoshi　Koumei Sugiura

*1中部大学　*2国立研究開発法人 情報通信研究機構
Chubu University　National Institute of Information and Communications Technology

Asynchronous Advantage Actor-Critic (A3C) is a representative method of deep reinforcement learning and is possible to solve difficult tasks such as games and robot control. However, it is difficult for deep reinforcement learning including A3C to understand and to explain the reason of action selection. To address this problem, we propose a method called a Mask-Attention A3C, which performs mask processing on feature map of Policy branch using attention map. The propose method can obtain an attention map that is useful for a visual explanation of agent behavior. In the experiment with Atari2600, we compare the scores in each game and demonstrate that the attention map obtained from our method is useful for visual explanation. In addition, we evaluate the explainability of obtained attention map using the scores of each game by changing the attention region.

1. はじめに

強化学習は，どのような行動を選択すれば良いかは教えず，どの行動を選択すればより良い結果に結び付くかを見つけ出す問題である．教師あり学習のように教師データを用意する必要がない点から，教師データの作成が困難であるタスクへ応用されている．例えば，ロボットの自律制御 [Gu 17] [Rajeswaran 17] やゲーム攻略 [Justesen 17] [Firoiu 17] である．ゲーム攻略において，Silver らのコンピュータ囲碁プログラム AlphaGo [Silver 16] がプロ囲碁棋士に勝利したことで注目された．また，2015 年に Q 学習と畳み込みニューラルネットワーク (CNN) を組み合わせた手法である Deep Q-Network (DQN) [Mnih 15] が提案され，Atari2600 において人間のプレイヤーを上回るスコアを達成した．DQN が登場して以降，強化学習に深層学習を取り込んだ深層強化学習が主流となり，画像のような状態数が膨大な問題を扱うことが可能となった．

しかし，深層強化学習は，エージェントが環境を探索することで学習データの収集を行う点や推論時におけるネットワーク内部の演算が複雑である点など，学習により獲得した行動の判断根拠が不明確である．そのため，ネットワークの推論結果である行動に対して判断根拠を解析することは，学習後のネットワークが十分信頼にできるか証明することや間違った行動をした場合に原因究明をする上で重要である．そこで，画像認識分野では，ネットワークの注視領域をヒートマップとして可視化した Attention map を用いることで，判断根拠の視覚的説明を実現している [Zhou 16] [Selvaraju 17] [Fukui 19]．

本研究では，深層強化学習の代表的な手法である Asynchronous Advantage Actor-Critic (A3C) [Mnih 16] に Attention 機構を導入した Mask-Attention A3C (Mask A3C) を提案する．推論時において，入力画像に対するネットワークの注視領域を可視化することで，学習によって獲得した行動の視覚的説明を実現する．

連絡先:
板谷 英典：itaya@mprg.cs.chubu.ac.jp
平川 翼：hirakawa@mprg.cs.chubu.ac.jp
山下 隆義：takayoshi@cs.chubu.ac.jp
藤吉 弘亘：hf@cs.chubu.ac.jp
杉浦 孔明：komei.sugiura@nict.go.jp

2. 関連研究

視覚的説明は様々なタスクで応用されており，ここでは画像認識，深層強化学習での主な手法を述べる．

2.1 画像認識における視覚的説明

画像認識分野におけるネットワークの推論結果に対する判断根拠の解析では，ネットワークが推論時に注視した領域をヒートマップとして可視化した Attention map を用いる手法がいくつか提案されている．Zhou らは，畳み込み層の応答値と全結合層の重みから，特定のクラスに応じた Attention map を獲得する Class Activation Mapping (CAM) [Zhou 16] を提案している．しかし，CAM は畳み込み層と全結合層間に Global Average Pooling (GAP) を導入する点など，ネットワーク構造の一部を変更する必要があるため，画像認識において性能が低下してしまう．そこで，Selvaraju らは，順伝播時の畳み込み層の応答値と逆伝播時の勾配を用いることで Attention map を獲得する Gradient-weighted CAM (Grad-CAM) [Selvaraju 17] を提案している．また，Attention map を用いた認識精度の向上も取り組まれている．福井らは，Attention map を Attention 機構へ応用した Attention Branch Network (ABN) [Fukui 19] を提案している．この手法では，Attention map による判断根拠の視覚的説明と認識精度の向上を同時に実現している．

2.2 深層強化学習および視覚的説明

深層強化学習の代表的な手法に，Asynchronous Advantage Actor-Critic (A3C) [Mnih 16] がある．A3C とは，複数環境における非同期的なパラメータ更新である Asynchronous，数ステップ先の報酬を考慮し学習する Advantage を導入した Actor-Critic 法ベースの手法である．学習に用いる経験の生成を並列実行し高速化することで，Atari2600 において短時間で高いスコアを達成している．

深層強化学習において，ネットワークの出力である行動の判断根拠を解析する研究はいくつか行われている．Sorokin らは，Value ベースの代表的な手法である Deep Q-Network に Attention 機構を導入した Deep Attention Recurrent Q-Network (DARQN) [Sorokin 15] を提案している．DARQN

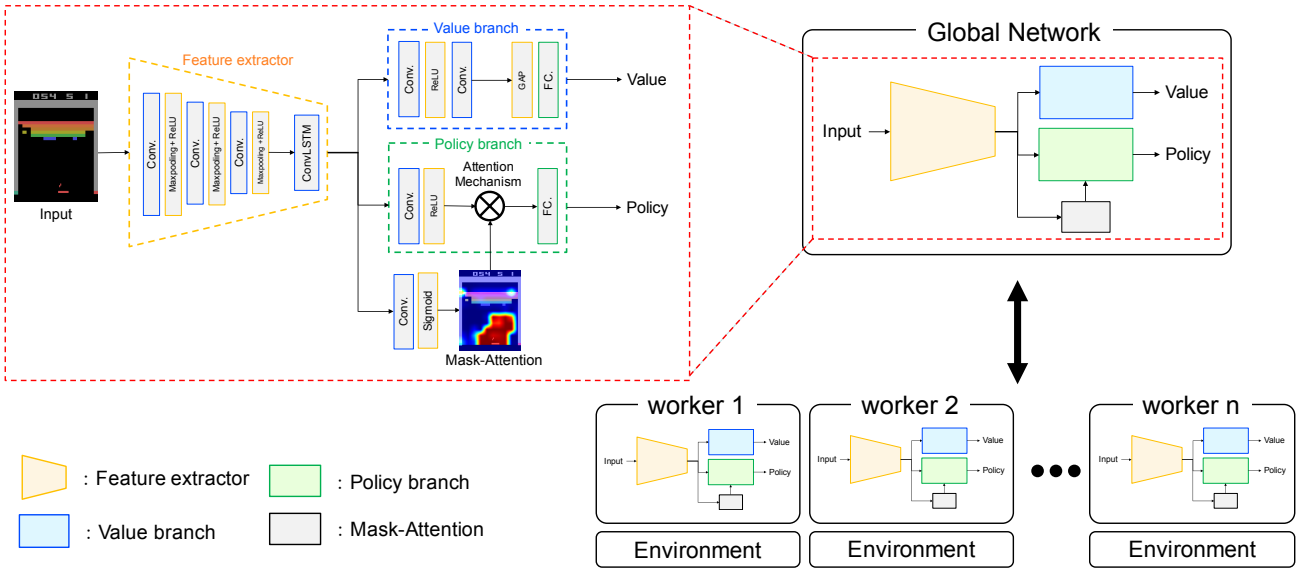


図 1: Mask A3C のネットワーク構造

は，行動価値関数 $Q(s,a)$ を出力するネットワークの中間層に Attention 機構を導入している．それに対し，提案手法では Policy 及び Value に基づいた A3C に Attention 機構を導入している．また，DARQN では Attention map の出力に softmax 関数を適用しているが，提案手法では sigmoid 関数を用いることで重要でないフレームは注視しない Attention map を獲得している．Greydanus ら [Greydanus 18] は，逆伝播時の勾配からガウシアンフィルターを適用した摂動画像を算出することで，A3C において Saliency map を獲得している．この手法は，Grad-CAM 同様に勾配を用いるため，Saliency map の獲得には逆伝播を行うする必要がある．それに対し，提案手法では，A3C のネットワークに Attention 機構を導入することで，逆伝播を行わず Attention map を獲得することができる．

3. Mask-Attention A3C

深層強化学習において，ネットワークの出力である行動の判断根拠を明確にすることは，学習後のネットワークが十分信頼に足るか証明することや選択した行動が間違っていた場合の原因究明をする上で重要である．そこで，行動の判断根拠を可視化するために，代表的な深層強化学習手法である A3C に Attention 機構を導入した Mask-Attention A3C (Mask A3C) を提案する．

3.1 ネットワーク構造

図 1 に提案する Mask A3C のネットワーク構造を示す．Mask A3C のネットワーク構造は Feature extractor，Value branch，Policy branch，Attention 機構，Mask-Attention から構成される．ここで，Feature extractor と Value branch，Policy branch は，従来の A3C におけるネットワーク構造と同様のものである．そのため，Value branch は状態価値関数を出力するブランチであり，Policy branch は方策を出力するブランチである．また，従来の A3C は Long-Short Term Memory (LSTM) を導入することで性能を大幅に向上している．しかし，Mask A3C に LSTM を導入すると入力画像に対する位置情報が欠落してしまうため，Attention map を出力することができない．そのため，位置情報を保持した LSTM である Convolutional LSTM (ConvLSTM) [Shi 15] を用いる．

3.2 Attention 機構

Mask A3C では，Policy branch に Attention 機構を導入することで，獲得した Attention map を考慮して方策を出力する．Attention 機構は，Policy branch 内における中間層の特徴マップに対し，Attention map を用いてマスク処理を行う．本論文ではマスク処理に用いる Attention map を Mask-Attention と呼称する．特徴マップに対する Mask-Attention を用いたマスク処理を式 (1) に示す．ここで，$F(\mathbf{s}_t)$ は Policy branch の中間層の特徴マップ，$M(\mathbf{s}_t)$ は Mask-Attention，$F'(\mathbf{s}_t)$ はマスク処理後の特徴マップである．

$$F'(\mathbf{s}_t) = F(\mathbf{s}_t) \cdot M(\mathbf{s}_t) \tag{1}$$

また，Mask-Attention は，Feature extractor から出力される特徴マップに対し，$1 \times 1 \times$ チャネル数 の畳み込み層と sigmoid 関数を適用することで獲得する．

4. 評価実験

提案手法である Mask A3C の有効性を確認するため，OpenAI gym [Brockman 16] のゲーム環境を用いて評価実験を行う．使用するゲームは，“Breakout”，“MsPacman”，“SpaceInvaders”の 3 種類である．比較対象は A3C と Mask A3C である．比較対象である A3C のネットワーク構造は，Mask A3C において Attention 機構と Mask-Attention ($1 \times 1 \times 1$ Conv.，sigmoid) を除いたネットワークとする．ネットワークへの入力情報はゲーム画面のグレースケール画像とし，出力は各ゲームにおける操作とする．学習条件は worker 数を 35，学習係数を 0.0001，割引率を 0.99 とする．学習終了条件は global step 数が 1.0×10^8 に到達した場合とする．また，エピソードの終了条件は各ゲームにおける 1 プレイ終了，及び step 数が $10,000$ に到達した場合とする．評価方法は以下の 3 通りである．

1. 各ゲーム環境でのスコア比較
2. Mask-Attention の注視領域を反転した場合のスコア比較
3. Mask-Attention を用いた行動の視覚的説明

表 1: 各ゲーム環境おける最大・平均スコア

	最大スコア		平均スコア	
	A3C	Mask A3C	A3C	Mask A3C
Breakout	864	864	613.86	584.30
MsPacman	4410	6670	3382.60	4207.90
SpaceInvaders	20130	20100	18998.40	19637.95

表 2: Mask Attention での注視領域の反転によるスコア比較

	最大スコア		平均スコア	
	normal	inverse	normal	inverse
Breakout	864	7	584.30	4.23
MsPacman	6670	630	4207.90	298.20
SpaceInvaders	20100	630	19637.95	292.65

4.1 スコア比較

A3C と Mask A3C での，各ゲーム環境における 100 エピソード間の最大・平均スコアを比較する．各ゲーム環境における最大・平均スコアを表 1 に示す．また，表 1 に示したスコアは，A3C と Mask A3C 共に 5 試行ずつ学習を行い，平均スコアが最も高かったモデルのスコアである．

表 1 から，Breakout における最大スコアは A3C と Mask A3C ともに 864 である．このスコアは，Breakout で獲得できる最高スコアである．それに対し，Breakout における平均スコアは A3C と比較し Mask A3C が 29.56 低いことが分かる．MsPacman における最大スコアは Mask A3C が A3C より 2260 高く，平均スコアも同様に 825.3 高いことが分かる．SpaceInvaders における最大スコアは A3C が 20130，Mask A3C が 20100 と同等のスコアであり，平均スコアは Mask A3C が A3C より 639.55 高いことが分かる．これらの結果から，Breakout は自身がボールを打ち返すのみで外的要因のない単純なタスクであるため，A3C においても高いスコアを獲得できていると考えられる．それに対し，MsPacman と SpaceInvaders では，敵などの外的要因を考慮して行動を選択する必要がある．そのため，Attention 機構の導入により様々な領域を注視する Mask A3C が A3C と比較し，高いスコアを獲得したと考えられる．

4.2 視覚的説明における Mask-Attention の有効性

Mask A3C による行動の視覚的説明を行うにあたり，Mask-Attention がネットワークの出力である行動に対して有効な注視領域を表しているか検証する．検証方法として，Mask-Attention の注視領域を反転させたマップを作成し，そのマップを Attention 機構に用いる場合でのスコアを算出する．そして，注視領域を反転させた場合と反転しない場合におけるスコアを比較することで，Mask-Attention が行動の視覚的説明に有効であることを確認する．Mask-Attention の注視領域を反転させたマップの作成を式 (2) に示す．ここで，$M(\mathbf{s}_t)$ は Mask-Attention，$M_{\text{inverse}}(\mathbf{s}_t)$ は Mask-Attention の注視領域を反転したマップである．

$$M_{\text{inverse}}(\mathbf{s}_t) = 1 - M(\mathbf{s}_t) \tag{2}$$

表 2 に，Mask-Attention における注視領域の反転によるスコア比較を示す．ここで，注視領域を反転しなかった場合が normal，注視領域を反転した場合が inverse である．

表 2 から，全ゲームにおいて inverse のスコアが normal と比較し著しく低下していることが分かる．したがって，Mask-Attention の注視領域は，高スコアを獲得する行動に対して有益な注視領域を表していると言える．

4.3 Mask-Attention を用いた行動の視覚的説明

各ゲームにおける Mask-Attention の可視化例を図 2 示す．Mask-Attention 下部のコントローラは，現フレームにおいてネットワークが推論した行動を表している．また，白矢印は Breakout ではボールの進行方向，MsPacman ではパックマンの進行方向，SpaceInvaders ではプレーヤーの攻撃であるビームの進行方向を示している．

図 2(a) Breakout から，Frame1,2 の白丸で示す通り，ブロック左端の反応が強くなっていることが分かる．Breakout はブロック上部へボールを通すと上部でバウンドを繰り返し大量にスコアを獲得できる．Frame1 から 2 では左側のブロックを崩しブロック上部へボールを通せるようになったため，ブロック左端の反応が強くなったと考えられる．また，Frame3,4 からボールをパドルで跳ね返す際にボールとボール周囲に強く反応し，ネットワークが出力する行動もボールの進行方向に合わせるように行動していることが分かる．ここから，ネットワークはボールを跳ね返す際のみにボールに着目し行動を選択していると考えられる．

図 2(b) MsPacman から，Frame1,2 において “上へ進む”行動を選択しているが，Frame3 以降では “左へ進む”行動を選択していることが分かる．また，Frame1,2 からパックマン周囲の注視領域に敵が入り込んでいないが，Frame3 でパックマン周囲の注視領域に敵 (白丸) が入り込んでいることが確認できる．ここから，ネットワークは Frame3 で敵を認識し，敵を回避するため “左側へ進む”行動を選択したと考えられる．

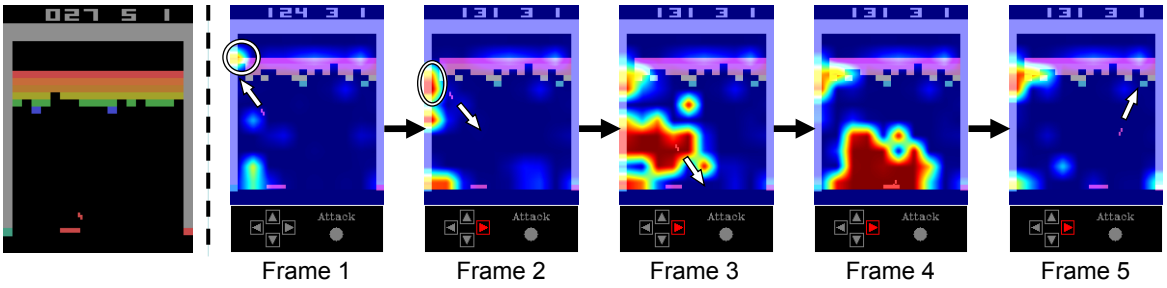
図 2(c) SpaceInvaders では，上部の UFO に Frame2 で弱く反応し，Frame5 で強く反応していることが確認できる．また，Frame2 での行動は “攻撃する”であり，攻撃であるビームは Frame5 において UFO に当たっていることが確認できる．SpaceInvaders において，UFO は一定周期で出現し撃破することで高得点を獲得できる．ここから，ネットワークは Frame2 で UFO を認識し撃破するための攻撃を行ったことで，Frame5 で UFO を撃破できたと考えられる．

5. おわりに

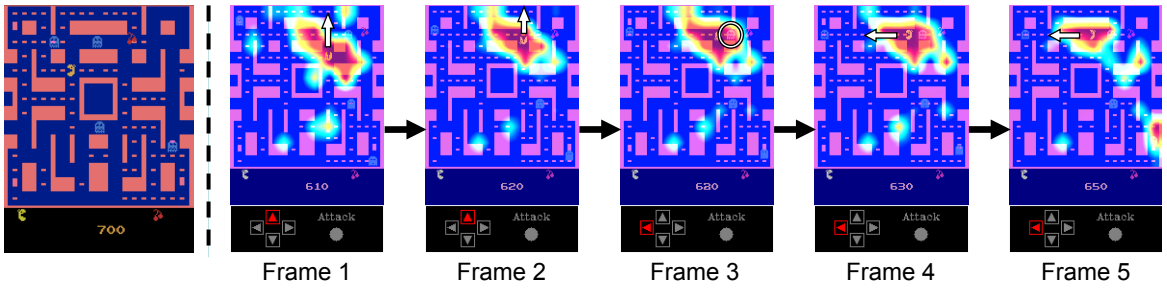
本研究では，深層強化学習の代表的な手法である A3C に Attention 機構を導入した Mask A3C を提案した．提案手法では，推論時に Mask-Attention を可視化することで，エージェントが選択する行動に対する判断根拠の視覚的説明を実現した．また，Atari2600 において，Mask-Attention を用いてエージェントの行動に対して視覚的説明を行い，各ゲームにおけるエージェントの有益な解析を示した．今後の課題としては，Value-branch に対して Mask-Attention を用いた Attention 機構を導入することで，状態価値関数に対する視覚的説明などが挙げられる．

謝辞

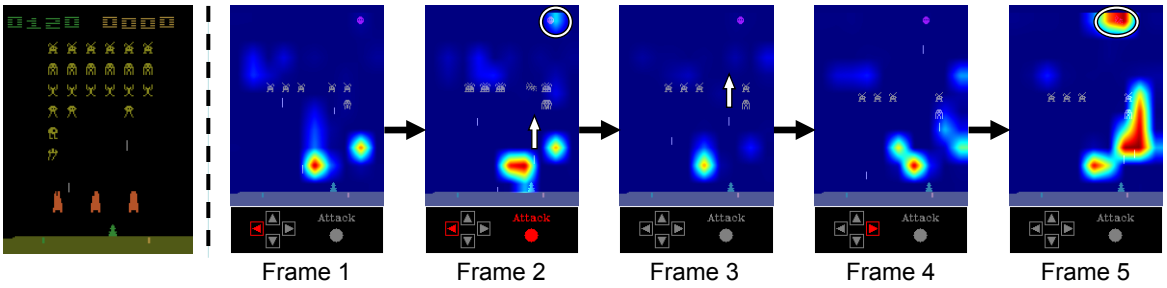
本研究は「次世代人工知能・ロボット中核技術開発」(人工知能の信頼性に関する技術開発) の支援を受けたものである．



(a) Breakout



(b) MsPacman



(c) SpaceInvaders

図 2: Mask-Attention の可視化例

参考文献

[Brockman 16] Brockman, G., *et al.*: OpenAI gym, *arXiv preprint, arXiv:1606.01540* (2016)

[Firoiu 17] Firoiu, V., *et al.*: Beating the World's Best at Super Smash Bros. Melee with Deep Reinforcement Learning, *arXiv preprint, arXiv:1702.06230* (2017)

[Fukui 19] Fukui, H., *et al.*: Attention Branch Network: Learning of Attention Mechanism for Visual Explanation, *CVPR* (2019)

[Greydanus 18] Greydanus, S., *et al.*: Visualizing and understanding Atari agents, *ICML*, Vol. 80, pp. 1792–1801 (2018)

[Gu 17] Gu, S., *et al.*: Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates, in *ICRA*, pp. 3389–339 (2017)

[Justesen 17] Justesen, N., *et al.*: Deep Learning for Video Game Playing, *arXiv preprint, arXiv:1708.07902* (2017)

[Mnih 15] Mnih, V., *et al.*: Human-level control through deep reinforcement learning, *Nature*, Vol. 518, No. 7540, pp. 529–533 (2015)

[Mnih 16] Mnih, V., *et al.*: Asynchronous Methods for Deep Reinforcement Learning, in *ICML*, pp. 1928–1937 (2016)

[Rajeswaran 17] Rajeswaran, A., *et al.*: Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations, *arXiv preprint, arXiv:1709.10087* (2017)

[Selvaraju 17] Selvaraju, R. R., *et al.*: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, *ICCV* (2017)

[Shi 15] Shi, X., *et al.*: Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, *NIPS*, pp. 802–810 (2015)

[Silver 16] Silver, D., *et al.*: Mastering the game of Go with deep neural networks and tree search, *Nature*, Vol. 529, No. 7587, p. 484 (2016)

[Sorokin 15] Sorokin, I. and Seleznev, A.: Deep attention recurrent q-network, *NIPS Workshop* (2015)

[Zhou 16] Zhou, B., *et al.*: Learning deep features for discriminative localization, *CVPR* (2016)