Attention 機構による視覚的説明を可能とした深層強化学習に関する研究

TP19001 板谷英典　　指導教授：山下隆義

1. はじめに

深層強化学習は，エージェントが環境とのインタラクションを通じて得る報酬を頼りに，状況に応じて最適な行動選択するための方策を学習する手法である．方策はネットワークで表現されており，ネットワーク内部の演算は複雑である．そのため，エージェントの行動選択に対して判断根拠を解析することは非常に困難である．

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) [1] に Attention 機構を導入した Mask-Attention A3C (Mask A3C) を提案する．Mask A3C では，方策と状態価値に対するエージェントの注視領域を表現した Mask-attention を獲得できる．推論時に Mask-attention を可視化することで，エージェントの行動選択に対する視覚的説明の実現を目的とする．また，Attention 機構の導入により，Mask-attention を考慮して学習することで，エージェントの性能向上を図る．

2. Asynchronous Advantage Actor-Critic

A3C [1] は，Asynchronous と Advantage を導入した Actor-Critic 法ベースの深層強化学習手法である．Asynchronous は複数環境における非同期でのパラメータ更新，Advantage は数ステップ先の報酬を考慮した学習である．A3C のネットワーク構造は，畳み込み層により特徴マップを抽出する Feature extractor，方策を出力する Policy branch，状態価値を出力する Value branch から構成される．方策はある状態において選択する行動の確率分布である．状態価値はある状態における報酬の期待値であり，ある状態にいることの価値を表す．

3. Mask-Attention A3C

エージェントの行動選択に対する判断根拠を解析するため，A3C に Attention 機構を導入した Mask-Attention A3C (Mask A3C) を提案する．Mask A3C では，Policy branch と Value branch に対し Attention 機構を導入することで，各ブランチの出力に対して注視した領域を表す Mask-attention を獲得する．推論時における各ブランチの Mask-attention を可視化することで，方策と状態価値の異なる 2 つの視点から，エージェントの行動選択に対する視覚的説明を実現する．

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 は，Feature extractor，Policy branch，Value branch，Attention 機構から構成される．従来の A3C では，時系列情報を考慮するため，Feature extractor に LSTM を用いる．しかし，Mask A3C に LSTM を導入すると，入力画像に対する空間情報が欠落するため，Mask-attention を獲得できない．そこで，時空間情報を考慮できる Convolutional LSTM (ConvLSTM) [2] を導入する．ブランチ i の Mask-attention M_i は，Feature extractor から出力される特徴マップ F_i に対し，$1 \times 1 \times$ # of chanels の畳み込み層と Sigmoid 関数を適用することで獲得する．

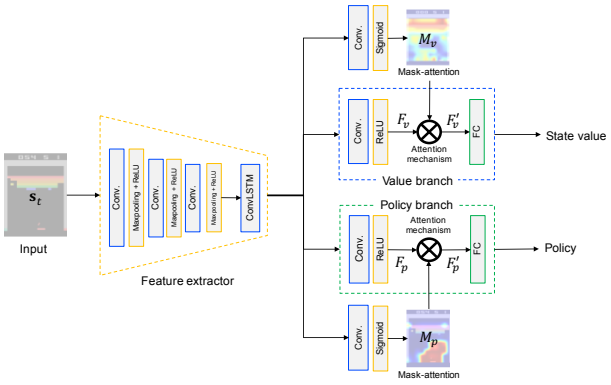


図 1: Mask A3C のネットワーク構造

3.2 Attention 機構

Mask A3C では，Policy branch と Value branch に Attention 機構を導入する．これにより，獲得した Mask-attention M_i を考慮し，方策及び状態価値を出力する．Attention 機構は，ブランチ i 内における中間層の特徴マップ F_i に対し，Mask-attention M_i を用いてマスク処理を行う．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，$\mathbf{s}_t$ は入力である現状態，$F_i(\mathbf{s}_t)$ はブランチ i 内における中間層の特徴マップ，$M_i(\mathbf{s}_t)$ はブランチ i における Mask-attention，$F'_i(\mathbf{s}_t)$ はマスク処理後の特徴マップである．

$$F'_i(\mathbf{s}_t) = F_i(\mathbf{s}_t) \cdot M_i(\mathbf{s}_t) \tag{1}$$

4. 評価実験

Mask A3C の有効性を確認するため，OpenAI Gym [3] のゲームタスクを用いて評価実験を行う．

4.1 実験概要

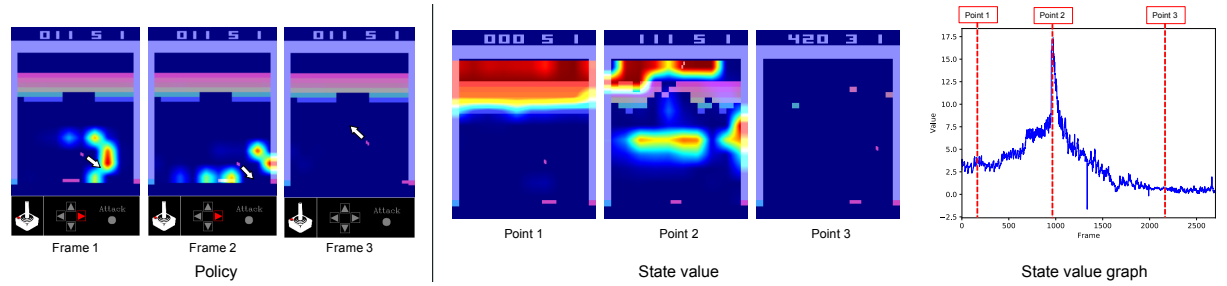
使用するゲームは，Breakout と Ms.Pac-Man，Space Invaders の 3 種類である．比較手法は，A3C と Mask A3C，各ブランチのみに対して Attention 機構を導入した Mask A3C (Policy Mask A3C，Value Mask A3C) の計 4 つである．入力はゲーム画面のグレースケール画像とし，出力である行動は各ゲームにおける操作とする．学習条件は worker 数を 35，学習係数を 0.0001，割引率を 0.99 とする．学習終了条件は global step 数が 1.0×10^8 に到達した場合とする．また，エピソードの終了条件は各ゲームにおける 1 プレイ終了，及び step 数が 1.0×10^4 に到達した場合とする．

4.2 スコア比較

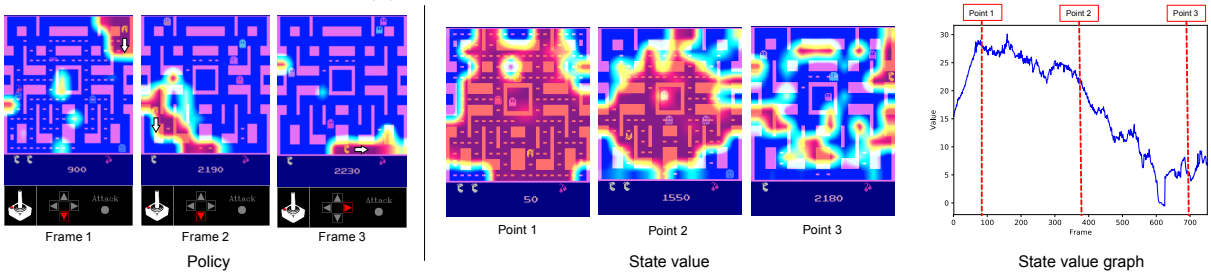
各ゲームタスクにおける 100 エピソード間の最大/平均スコアを表 1 に示す．表 1 から，Breakout における最大スコアは全手法において 864 である．このスコアは，Breakout で獲得できる最高スコアである．また，Breakout における平均スコアは，Mask A3C が A3C と比較しほぼ同等である．Ms.Pac-Man では，最大/平均スコア共に Mask A3C が最も高いスコアである．Space Invaders では，最大スコアは Policy Mask A3C，平均スコアは Mask A3C が最も高いスコアである．Breakout はパドルでボールを打ち返すのみであり，外的要因のない単純なタスクである．そのため，A3C と Mask A3C が同等のスコアであったと考えられる．一方，Ms.Pac-Man と Space Invaders 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．Policy branch に Attention 機構を導入した Policy Mask A3C と Mask A3C では，クッキーやインベーダーなど，方策に関連した領域を強調する．そのため，A3C と比較しスコアが向上したと考えられる．

4.3 Mask-attention を用いた視覚的説明

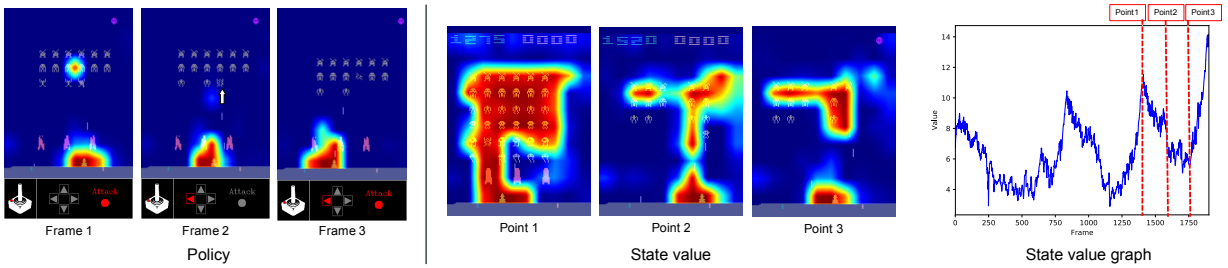
各ゲームにおける Mask-attention の可視化例を図 2 に示す．図 2(a) の Policy から，Frame 1 ではボールの進行方向を注視している．Frame 2 ではパドルを含めたボール周囲を注視し，ボールを打ち返した後である Frame 3 では注視している領域がない．ここから，Breakout ではボールを打ち返す直前において，ボールの進行方向を予測しパドルを制御していると考えられる．図 2(b) の Policy から，Frame 1 ではパックマン周囲を注視している．Frame 2 では画面内に存在するクッキーを注視し，Frame 3 では Frame 2 における注視領域にパックマンが移動している．ここから，Ms.Pac-Man ではパックマンの周囲を警戒しながら，クッキーへ向かうように制御していると考えられる．図 2(c) の Policy から，Frame 1 でインベーダーを注視し，行動は攻撃を選択している．そして，Frame 3



(a) Breakout: 白矢印はボールの進行方向を示す.



(b) Ms.Pac-Man: 白矢印はパックマンの進行方向を示す.



(c) Space Invaders: 白矢印はビームの進行方向を示す.

図 2: Mask-attention の可視化例: Policy における下部のコントローラは，現フレームでモデルが推論した行動である.

表 1: 各ゲームおける 100 エピソード間の最大/平均スコア: 各手法で 5 試行ずつ学習し，平均スコアが最も高いモデルのスコアを示す．max/mean は最大/平均スコアである.

	Att. mechanism		Breakout		Ms.Pac-Man		Space Invaders	
	Policy	Value	max	mean	max	mean	max	mean
A3C			864	**662.0**	5380	4573.3	19505	18531.8
提案手法	✓		864	595.8	6330	4833.8	19860	19102.8
		✓	864	606.9	4830	4044.5	19675	18537.8
	✓	✓	864	640.0	6610	**5314.1**	19810	**19212.5**

表 2: Mask-attention の反転によるスコア比較: normal は Mask-attention を反転しない場合，inverse は反転した場合，random はランダムに行動選択した場合である.

Att. mechanism			Breakout		Ms.Pac-Man		Space Invaders	
Policy	Value		max	mean	max	mean	max	mean
✓		normal	864	595.8	6630	4833.8	19860	19102.8
		inverse	4	2.2	290	268.9	805	306.9
✓	✓	normal	864	640.0	6610	5314.1	19810	19212.5
		inverse	5	1.8	410	194.4	915	420.2
random			5	1.2	1080	247.8	460	142.1

において Frame 1 で注視したインベーダーを撃破している．ここから，Space Invaders では撃破するインベーダーを認識し，エージェントを制御していると考えられる．また，図 2(a)(b)(c) の State value から，Breakout ではブロック，Ms.Pac-Man ではクッキー，Space Invaders ではインベーダーを注視している．そして，注視した物体の減少に合わせ注視領域が縮小している．これらの結果から，Policy の Mask-attention は現状態に対し行動に直結する物体，State value の Mask-attention はスコアに寄与する物体を表していると考えられる.

4.4 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 がモデルの出力である方策に対して有益な領域を表しているか検証する．検証方法として，Mask A3C における Policy branch の Mask-attention を反転したマップを作成し，そのマップを Attention 機構に用いた場合のスコアを算出する．Mask-attention を反転する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．表 2 に，Mask-attention の反転によるスコア比較を示す．表 2 から，全ゲームにおいて inverse が normal と比較して著しくスコアが低下し，random と同等のスコアである．したがって，Policy branch における Mask-attention は，高スコアを獲得する行動に対して有益な領域を獲得したと言える.

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．Mask A3C では，推論時に Mask-attention を可視化することで，エージェントの行動選択に対する判断根拠の視覚的説明を実現した．今後は，ロボット制御などの更に複雑なタスクへの適用と可視化に取り組む予定である.

参考文献

[1] V. Minh, *et al.*, “Asynchronous Methods for Deep Reinforcement Learning”, ICML, 2016.

[2] X. Shi, *et al.*, “Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting”, NeurIPS, 2015.

[3] G. Brockman, *et al.*, “OpenAI Gym”, *arXiv preprint arXiv:1606.01540*, 2016.

研究業績

[1] 板谷英典 等, “A3C における Attention 機構を用いた視覚的説明”, 人工知能学会全国大会, 2020. (学生奨励賞受賞)
(他 2 件)