Mask-A3C による強化学習における方策の視覚的説明の解析

EP16028 尾関菜津美　　指導教授：山下隆義

1. はじめに

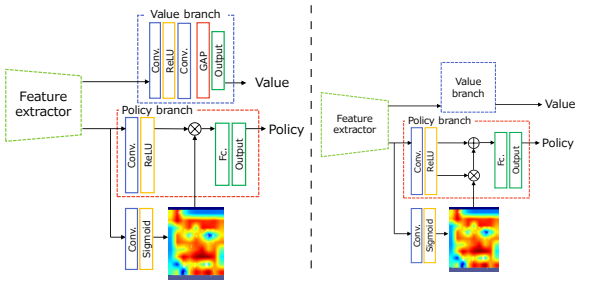
強化学習は，ある環境における報酬を最大化する行動をエージェントが学習するタスクである．代表的な強化学習法である Asynchronous Advantage Actor-Critic(A3C)[1] は，異なる環境を持つ複数の worker を非同期に並列学習することで効率的かつ高速に学習できる．しかし，強化学習により獲得した行動の判断根拠は不明確であるという問題点がある．本研究では，A3C の Policy branch に対し，Attention map を用いてマスク処理を行う Mask Attention A3C(Mask-A3C) を提案する．Mask Attention を用いることで，獲得した方策の判断根拠に対する視覚的説明を獲得することが可能となる．

2. A3C

Asynchronous Advantage Actor-Critic(A3C)[1] は，3 つのアイデアを組み合わせた深層強化学習手法である．1 つ目は，非同期な複数環境で学習をする Asynchronous である．2 つ目は，2 ステップ以上先の報酬まで考慮した学習をする Advantage である．3 つ目は，エージェントの行動と状態価値を独立して学習する Actor-Critic である．これらの組み合わせにより，非同期な分散並列学習を効率的に行うことができる．

3. 提案手法

行動の判断根拠を可視化するために，A3C における Policy branch の中間層にマスク機構を導入した Mask-A3C を提案する．図 1(a) に Mask-A3C のネットワーク構造を示す．Feature extractor は，3 層の畳み込み層と ConvLSTM により入力画像から特徴マップを抽出する．Value branch は，Feature extractor で抽出した特徴マップから状態価値を出力する．Policy branch は，Feature extractor で抽出した特徴マップと Attention map から方策を出力する．本研究ではマスク機構について，残差を導入した残差あり Mask-A3C も提案する．図 1(b) に残差を導入したネットワークの構造を示す．



(a) Mask-A3C　　(b) 残差あり Mask-A3C

図 1: 提案手法のネットワーク構造

4. 実験概要

本実験では，従来手法である A3C との性能比較，および方策の判断根拠に対する視覚的説明の解析を行う．学習環境として OpenAI Gym[2] で提供されている Breakout, SpaceInvaders, MsPacman を用いる．学習条件は，worker 数を 35，global step を 1.0×10^8 とする．評価指標には，各ネットワークの 100 エピソードの平均スコアと最大スコアを用いる．

4.1. 実験結果

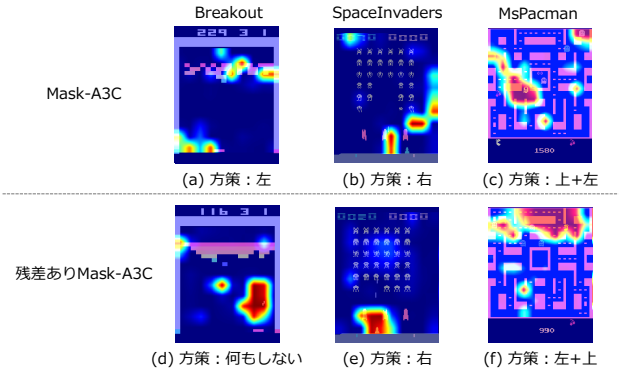
表 1 に各ゲームにおける平均と最大スコアを示す．表 1 から，Breakout では全てのネットワーク構造で最大スコアとなる 864 を獲得した．そのため，Breakout はマスク機構が効果を発揮しにくいゲームであると考える．SpaceInvaders では，A3C より Mask-A3C と残差あり Mask-A3C の方が高い平均および最大スコアを獲得した．A3C と比べ，Mask-A3C の平均スコアは約 3000 高いため，SpaceInvaders ではマスク機構が有効であると考えられる．一方，MsPacman では最大スコアは A3C，平均スコアは残差あり Mask-A3C が高いスコアを獲得した．以上のことから，一部のゲームにおいてマスク機構が有効であると言える．

表 1: 平均スコアと最大スコアの比較

		A3C	提案手法 (Mask-A3C) 残差なし	残差あり
mean (max)	Breakout	**613.86** (**864**)	466.98 (**864**)	504.37 (**864**)
	SpaceInvaders	16373.30 (19640)	**19269.45** (20160)	18348.45 (**21005**)
	MsPacman	**3382.60** (4410)	1814.20 (2490)	3066.30 (**4580**)

4.2. Attention map による視覚的説明

図 2 に各ネットワークが獲得した Attention map を示す．Breakout では各ネットワークともエージェントの手元に戻ってきた弾に反応している．さらに，Mask-A3C はエージェント自身にも反応している．図 2(a) のように，エージェントが左右どちらかに反応した場合，反応した方に移動する方策となった．また，ブロックの両端にも反応し，弾が壁の上部に入ると弾が当たったとき，ブロックにも反応している．SpaceInvaders では，残差あり Mask-A3C はエージェント自身に反応している．一方，Mask-A3C はエージェントに反応せず，エージェントと敵の間の空間に反応している．さらに，各ネットワークとも敵の数が少なくなると，エージェントの真上付近にいる敵にも反応する．MsPacman では，エージェントの進行方向に反応しており，残差あり Mask-A3C は Mask-A3C より広範囲に反応している．図 2(f) に示すように，エージェントの進行方向に敵がいる場合がある．この時，敵を避けるような方策となった．このことから，残差あり Mask-A3C は Mask-A3C より広い範囲を注視することで，先の行動の選択肢を増やしていると考えられる．全ゲームで Mask-A3C より残差あり Mask-A3C の方がより広い範囲に反応している．また，全ゲームでエージェント以外のスコアに関係ある場所にも反応している．これはエージェントがより良い報酬を得る行動を取るのに必要な情報を得るためだと考える．以上のことから，提案手法により行動の判断根拠を獲得できたのではないかと考える．



(a) 方策：左　(b) 方策：右　(c) 方策：上+左

(d) 方策：何もしない　(e) 方策：右　(f) 方策：左+上

図 2: 各ネットワークの Attention map

5. おわりに

本研究では，A3C の Policy branch に対して Attention map を用いてマスク処理を行う Mask-A3C を提案し，Mask Attention を用いて獲得した方策の判断根拠に対する視覚的説明を実現した．今後は他ゲームでも解析を行う．

参考文献

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

[2] G. Brockman, *et al.*, “OpenAI Gym”, arXiv, 2016.