# Critic-Attention による探索基準を用いた大規模環境における効率的な深層強化学習

EP14133 村瀬卓也　　指導教授：山下隆義

## 1. はじめに

深層強化学習は，エージェントが未知の環境において試行錯誤を行い，獲得した報酬を頼りに最適な行動を学習する手法である．エージェントは報酬に辿り着くまで，探索基準が無い状態でランダムに行動をする．そのため，報酬を獲得する機会が少ない大規模で複雑な環境においては，適切な行動を獲得するまでに膨大な試行回数を必要とする．そこで本研究では，アテンションマップを探索基準とすることで，深層強化学習を効率化する手法を提案する．提案手法では，エージェントを階層化して模倣学習することで，より効率的な探索の実現が期待できる．

## 2. Mask-Attention 機構

Mask-Attention 機構は，Mask-Attention A3C (Mask-A3C) [1] で用いられた，判断根拠の視覚的説明を行う機構である．中間層の特徴マップに対し，$1 \times 1 \times$ チャネル数 の畳み込み層と Sigmoid 関数を適用することで，アテンションマップを算出する．算出したアテンションマップを用いて，中間層の特徴マップに対し，マスク処理を行う．これにより，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ができる．

## 3. 提案手法

本研究では，アテンションマップによる探索基準を用いることで，強化学習を効率化する手法を提案する．図 1 に，提案手法の構造を示す．Mask-Attention 機構を導入した Critic モデルから Critic-Attention を獲得し，Policy の探索基準とする．それにより，学習の効率化を図る．また，Policy を階層化し，模倣学習を行う．以下に，提案手法の学習手順を示す．

STEP1：Critic モデルの事前学習
: 人間のプレイデータを用いて，Critic モデルの事前学習を行う．学習したモデルから，報酬に繋がるオブジェクトを注視した Critic-Attention を獲得する．

STEP2：Policy モデルの模倣学習
: STEP1 で獲得した Critic-Attention を，各 Policy モデルに入力することにより，報酬に繋がるオブジェクトを考慮しながら学習する．また，タスクを複数のサブタスクに細分化し，各 Sub Policy に割り当てる．Master policy は，ある状態がどのサブタスクに属するか学習する．そして Sub policy は，割り当てられたサブタスクに対して最適な行動を学習する．

STEP3：強化学習
: Critic モデルと各 Policy モデルは，A3C による強化学習を行う．Critic モデルと Master policy は，獲得した合計報酬から学習する．Sub policy は，自身が選ばれている時にのみ獲得した報酬で学習する．

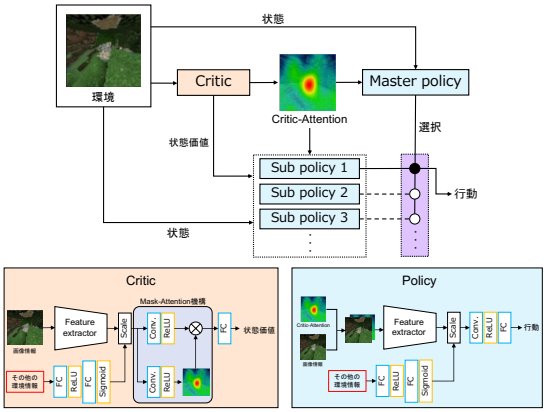


図 1: 提案手法の構造

## 4. 評価実験

提案手法の有効性を評価実験により示す．

### 4.1. 実験環境

本研究では，Minecraft を用いて強化学習を実験するために開発されたシミュレータである MineRL を用いる．MineRL では，人間のプレイデータが公開されているため，強化学習と模倣学習を組み合わせた実験ができる．最終目標はダイアモンドの獲得であり，目標を達成するまでに多くの工程が必要となる．報酬は，目標達成までに必要な中間アイテムを獲得することで獲得できる．図 2 に，MineRL の報酬設計を示す．最初の中間アイテムである原木から目標の達成条件であるダイヤモンドに近づくにつれて，獲得できる報酬値が増加する．

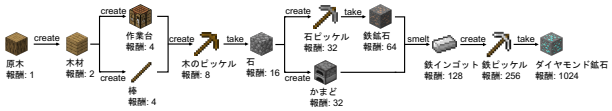


図 2: MineRL の報酬設計

### 4.2. 実験概要

MineRL に対して学習を行い，獲得した累積報酬のスコアを比較する．テスト回数は 100 エピソードとし，その平均値を最終スコアとする．また，学習時は環境とのインタラクション回数に，800 万ステップの上限を設ける．限られたステップ数で獲得したスコアを比較することで，効率的な探索が行えているか確認する．比較手法は，代表的な強化学習手法である A3C，A3C に模倣学習を追加した A3C+模倣学習，代表的な階層型強化学習である MLSH，MLSH に模倣学習を追加した MLSH+模倣学習，The MineRL Competition 2020 のトップ 3 チーム，提案手法の計 8 つである．ここで提案手法と MLSH は，Sub policy の数を 10 とする．

### 4.3. 実験結果

表 1 に，累積報酬の平均スコアを示す．表中の w/o は，模倣学習の有無を表す．表 1 から，提案手法が最も高いスコアを獲得していることがわかる．また，The MineRL Competition 2020 の 1 位と比較して，3.68pt 高いスコアを獲得した．一方で，A3C が最も低いスコアとなっている．これは，探索基準が無い強化学習では，なかなか報酬に辿り着けず，学習が進まないためである．また，MLSH でもほとんどスコアの向上が見られないため，強化学習のみの場合，階層化だけではスコアの向上に寄与しないことがわかる．しかし，A3C，MLSH ともに，模倣学習を追加した場合にスコアが向上している．特に MLSH+模倣学習の場合，通常の MLSH と比べてスコアが 32.35pt 向上していることから，階層型強化学習に模倣学習を行うことは有効であるといえる．提案手法では，さらに Critic-Attention を探索基準として用いることで，MLSH+模倣学習と比較して，スコアが 5.68pt 向上した．このことから，提案手法は効率的な探索が可能であるといえる．

表 1: 累積報酬の平均スコア

| 手法名 | A3C | | MLSH | | コンペティション | | | 提案手法 |
|---|---|---|---|---|---|---|---|---|
| | w/ | w/o | w/ | w/o | 3 位 | 2 位 | 1 位 | |
| スコア | 0.32 | 5.2 | 0.56 | 37.55 | 12.79 | 13.29 | 39.55 | **42.23** |

## 5. おわりに

本研究では，Critic-Attention を探索基準とすることで，効率的な探索を可能とする手法を提案した．MineRL を用いた実験により，提案手法の有効性を示した．今後の予定としては，Minecraft のさらなる攻略や，より大規模で複雑な環境への適用が挙げられる．

## 参考文献

[1] H. Itaya, *et al.*, “Visual Explanation using Attention Mechanism in Actor-Critic-based Deep Reinforcement Learn”, IJCNN, 2021.