# Mask-attention 機構を導入した PPO による物体把持動作の視覚的説明

○本多航也　板谷英典　平川翼　山下隆義　藤吉弘亘（中部大学）

## 1. はじめに

大型の物流倉庫や，生産現場などでは人員削減の観点からロボットマニピュレーションの需要が増加している．物流倉庫には様々な種類，形状の品種があり，その品種に合わせた把持をする必要がある．物体把持におけるロボットマニピュレーションでは，ロボット動作を操縦者が教示するティーチングプレイバックを用いるのが一般的である．ロボットの動作環境が異なると，再度操縦者によるティーチングが必要となるため，操縦者の負担は大きい．また，ティーチング操作はロボットにとって最適な判断であるとは限らない．この問題を解決するアプローチとして，深層強化学習 [1, 2, 3, 10, 11] を用いて最適なロボット動作を獲得する手法 [4, 5, 6, 7] が提案され，バラ積み物体から把持対象物体が直接ピッキングできない場合は，周りの物体を退避させるという動作の定義が可能となっている．しかしながら，従来の深層強化学習を用いたロボット動作はどのような判断根拠によって動作を定義しているのか不明である．

本研究では，画像入力の深層強化学習モデルに Mask-attention 機構を導入し，ロボット動作に対する判断根拠の視覚的説明を実現する．判断根拠の可視化は信頼性の向上に繋がり，Human-Robot Interaction の観点から重要である．タスクとして，複数のオブジェクトが存在する環境において把持対象物体を把持するタスクを対象とする．実験では，画像を用いた深層強化学習手法と提案手法を比較することで有効性を確認する．また，獲得した Mask-attention を可視化することでロボット動作における判断根拠を解析する．

## 2. 関連研究

深層強化学習モデルの出力である行動の判断根拠を視覚的に解析するアプローチには，摂動ベースの手法とアテンションベースの手法の 2 つがある．

### 2.1 摂動ベースの視覚的説明

摂動ベースの視覚的説明手法は，逆伝播時の勾配情報を用いて判断根拠の解析を行う．摂動ベースの手法として，Greydanus らは [8]，エージェントの行動選択を解釈するための Saliency map を生成する顕著性算出手法を提案している．入力画像に対し，特定の画素を欠落させる摂動を適応し摂動画像を作成する．その摂動画像を用いて，方策を出力することで，欠落させた画素の重要度を算出し Saliency map を生成する．

### 2.2 アテンションベースの視覚的説明

アテンションベースの視覚的説明手法は，勾配情報を必要とせず，順伝播時に判断根拠を獲得し解析する．アテンションベースの手法として，板谷ら [9] は，Asynchronous Advantage Actor-Critic (A3C) [10] に Mask-attention 機構を導入した Mask-Attention A3C (Mask A3C) を提案している．A3C は，複数環境においてパラメータを非同期に更新する Asynchronous と，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 手法である．Mask A3C は，Feature extractor，Policy branch，Value branch で構成される．Policy branch，Value branch に Mask-attention 機構をそれぞれ導入することで，対応するアテンションマップである Mask-attention を獲得できる．Mask A3C のネットワーク構造を図 1 に示す．Feature extractor は，複数の畳み込み層を用いて，入力画像に対して特徴抽出を行っている．また，ConvLSTM を導入することで入力画像に対する時空間情報を考慮している．Value branch と Policy branch は，それぞれ状態価値と方策を出力しているブランチである．これらのブランチは，Feature extractor の応答値を入力とし，各ブランチの中間層に Attention 機構を導入している．

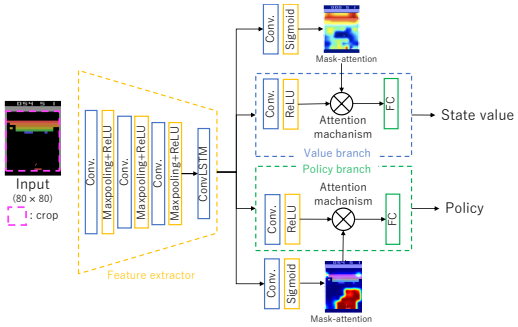


図 1　Mask A3C のネットワーク構造

Attention 機構では，獲得した Mask-attention を用いて各ブランチの中間層における特徴マップに対しマスク処理を施す．Mask-attention は，Feature extractor から出力される特徴マップに 1×1 の畳み込み層と Sigmoid 関数を適応し獲得する．Mask-attention を特徴マップで要素積をとり，マスク処理後の特徴マップを算出する．獲得した特徴マップに対し全結合層を用いて状態価値と方策を算出する．

## 3. 提案手法

ロボットマニピュレーションをティーチングする場合，人との協働を視野に入れて学習する必要がある．しかし，従来の深層強化学習では，ロボットの制御値を出力するネットワーク内部の演算が複雑である．そのため，ロボット動作に対する判断根拠が不明確であり，人がロボットの行動を理解ことが困難である．これは，人とロボットが協働するには望ましくない危険な状態である．そこで，画像入力の深層強化学習を用いたロボットマニピュレーションに Mask-attention 機構を導

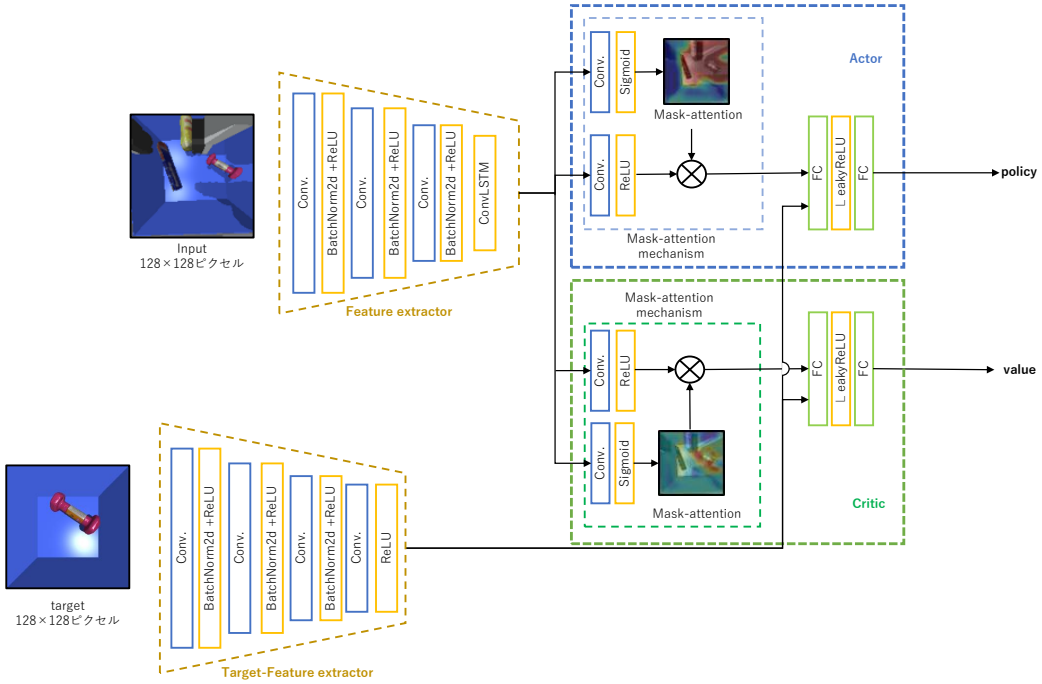


図 2　提案手法のネットワークモデル

入し，判断根拠を解析可能とするネットワークモデルを提案する．

### 3.1　提案手法の構造

提案手法のネットワークモデルを図 2 に示す．提案手法は，強化学習手法の Proximal Policy Optimization (PPO) [11] を用いる．PPO はクリッピングした代理目的関数とクリッピングしていない代理目的関数を比較し，より値の小さい関数を選択することで過度な方策更新を防ぐ手法である．Actor-Critic 構造を用いており，Actor で policy，Critic で value を算出する．本手法では，Actor と Critic それぞれに Mask-attention 機構を導入する．Mask-attention 機構は，Feature extractor から出力される特徴マップに畳み込み層と Sigmoid 関数を適用し，各ブランチの Mask-attention を獲得する．獲得した Mask-attention は最適な方策と状態価値に寄与する領域を強調し学習する．Mask-attention 機構を導入することで，policy，value に対応した Mask-attention を獲得できる．

実験環境には把持対象物体以外のオブジェクトも配置する．そのため，把持対象物体の情報を考慮した学習が必要である．そこで，把持対象物体用の Feature extractor である Target-Feature extractor で特徴量抽出を行い，各ブランチの全結合層に入力する．この構造により，複数のオブジェクトがある環境において把持対象物体を把持するための行動を獲得する．

### 3.2　提案手法の流れ

提案手法を用いてエージェントを制御し，学習を行う．提案手法の流れを以下に示す．

1. 環境にエージェントを作成し，提案手法のネットワークを作成する．環境で 1 ステップごとの状態を観測し，128 × 128 ピクセルの画像として，Feature extractor に入力する．
2. Feature extractor で算出した特徴マップを Actor と Critic それぞれに入力し，方策と状態価値に対応した特徴マップを算出する．同じ特徴マップを用いて，Actor と Critic それぞれの Mask-attention 機構で，方策と状態価値に対応した Mask-attention を算出する．算出した Mask-attention を用いて特徴マップにマスク処理を施す．
3. Target-Feature extractor に把持対象物体の 128 × 128 ピクセルの画像を入力し，把持対象物体に対する特徴マップを算出する．
4. マスク処理を施した特徴マップと把持対象物体に対する特徴マップを各ブランチの全結合層に入力し，方策と状態価値を算出する．その後，各エージェントで蓄積した報酬を用いて対応するネットワークの逆伝播を行う．

## 4.　評価実験

提案手法の有効性を確認するため，把持対象物体の把持タスクにおける収益と，タスク達成率による比較を行う．また，獲得した Mask-attention を用いた視覚的説明を行う．

### 4.1　実験環境

本研究では，Isaac-gym [13] を用いて作成したロボットマニピュレーションタスクを対象とする．Isaac-gym は NVIDIA が開発した強化学習研究のための物理シミュレーション環境である．シミュレーション環境と強化学習を GPU のみで実行でき，単一のワークステーションで数千の環境を並列に実行することができる．実験環境を図 3 に示す．本実験におけるタスクは，単腕アームロボットの Franka Emika 社の Panda を用いて複数のオブジェクトの中から把持対象物体を把持するタスクとする．把持するオブジェクトは ARC2017 RGB-D Dataset [12] にある 3 種類を用いる．使用するオブジェクトを表 1 に示す．表 1 のオブジェクトをトレイ

表 1　使用するオブジェクト

| 物体 | ダンベル | ラップ | テニスボール |
|---|---|---|---|
| 外見 |  |  |  |

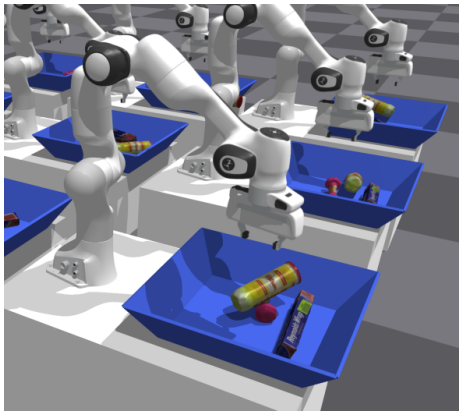

図 3　実験環境

にそれぞれ 1 つずつ出現させる．エージェントであるエンドエフェクタは，エピソード開始と同時に自動的に下降し，12 ステップでトレイに到達する．12 ステップ経過時にエンドエフェクタの開閉が自動的に行われ，その後エンドエフェクタの上昇を行う．本タスクは 12 ステップ間で把持対象物体を捉えることが目的である．エージェントが取れる行動は，トレイに対して奥か手前，トレイに対して左か右，エンドエフェクタの正回転か逆回転の 3 種類である．これらの行動は，エンドエフェクタの下降時のみ実行する．報酬は把持対象物体を把持し，一定以上持ち上げた場合のみ+1 とする．

### 4.2　評価実験

Mask-attention 機構を用いないネットワークモデルと提案手法である Mask-attention 機構を用いたネットワークモデルのスコア推移を比較する．学習は 25,000,000 ステップ行い，ステップごとのスコアで比較する．Waker 数は 128 で並列に学習する．把持対象物体はダンベルとする．タスクを達成したか確認するため，最適モデルで 60 回テストしタスク達成率を求める．最適モデルとは学習を 5 回試行し，スコアが最も高かったモデルである．タスクの達成は把持対象物体を把持できた場合とし，それ以外はタスク未達成とする．また，テスト時の Mask-attention を可視化することでロボット動作に対する視覚的説明を行う．獲得した Mask-attention が，有効な領域を示すか確認するため，Mask-attention を反転させた場合と反転していない場合のタスク達成率を算出し，各タスクの達成率を比較する．

### 4.3　タスク達成率による比較

#### スコア比較

Mask-attention 機構なしモデルと提案手法のエピソード毎の平均スコアの推移を図 4 に示す．図 4 から，提案手法は Mask-attention 機構なしモデルと比較して，同程度のスコアを獲得している．このことから，

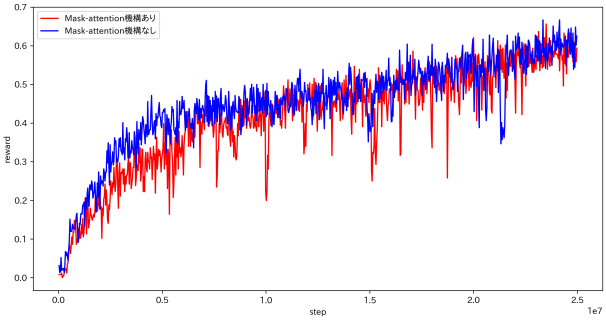


図 4　学習時における平均スコアの推移（5 回試行）

表 2　タスク達成率

| Mask-attention 機構 | なし | あり |
|---|---|---|
| 達成率 [%] | 51.67 | 51.67 |

ロボットマニピュレーションタスクにおいて提案手法が学習に悪影響を与えずに，Mask-attention を獲得可能であることが確認できる．

#### タスク達成率による比較

各手法におけるタスク達成率を表 2 に示す．表 2 から，提案手法のタスク達成率が 51.67pt，Mask-attention 機構なしモデルのタスク達成率が 51.67pt であり，同値であることがわかる．このことから，Mask-attention 機構の導入は，最適なロボット動作の獲得に悪影響がないといえる．

### 4.4　Mask-attention 機構の効果

#### Mask-attention の可視化

獲得した Mask-attention を時刻 $t$ ごとに，図 5 に示す．図 5 時刻 $t = 1$ における Value の Mask-attention では，ラップで半分以上隠れているダンベルにのみ注視していることが分かる．したがって，ダンベルを把持対象物体だと認識していると考えられる．一方，Policy の Mask-attention では，ダンベルの周りを強く注視していることがわかる．このことから，Mask-attention 機構によりダンベルが把持対象物体であると正しく認識し，把持できる場所を注視しているといえる．また，把持対象物体以外のオブジェクトの周囲にも注視しているが，オブジェクト自体を注視していない．このことから，把持対象物体以外のオブジェクトの位置のみ把握し，オブジェクト自体を注視しないことで把持しないようにしていると考えられる．時刻 $t = 7$ における Value の Mask-attention では，ダンベルの特徴的な部分を注視していることが分かる．このことから，エンドエフェクタがオブジェクトに近づき，オブジェクトの外形が分かるようになると，把持対象物体と似ているオブジェクトに注視していると考えられる．一方，Policy の Mask-attention では，ラップに対し注視せず，ダンベルの周りにのみ注視していることが分かる．このことから，オブジェクトとオブジェクトの境界線を認識し，ダンベルを把持できる部分を認識していると考えられる．

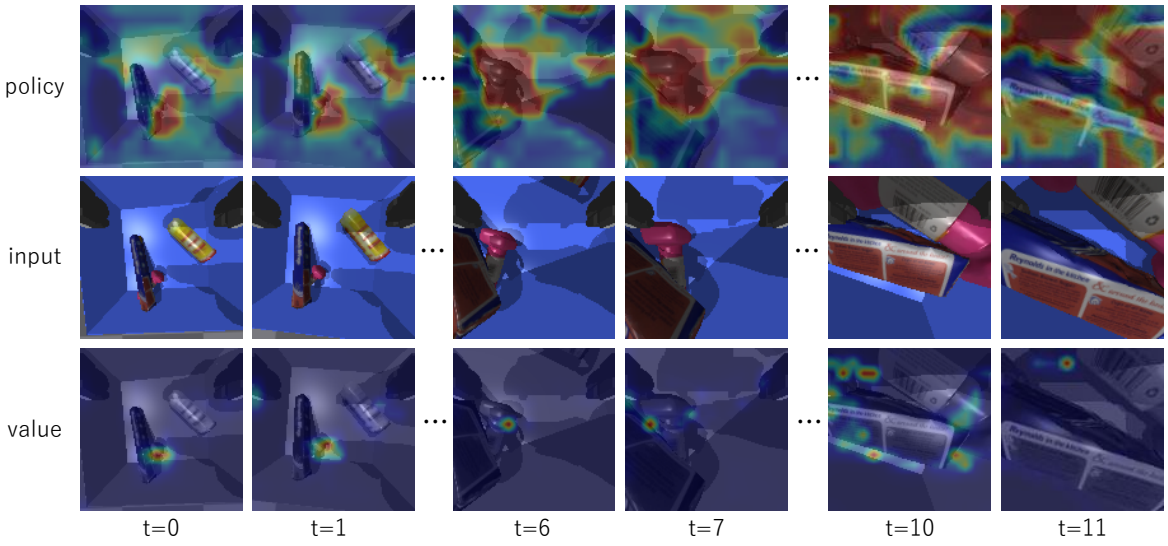


図 5　Mask-attention の可視化結果

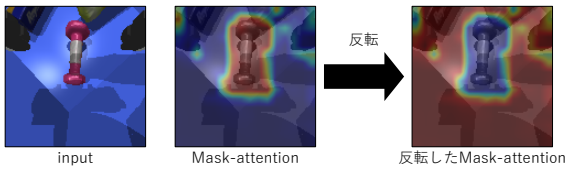


図 6　Mask-attention の反転例

表 3　Mask-attention の反転によるタスク達成率

| Mask-attention の反転 | なし | あり |
|---|---|---|
| 達成率 [%] | 51.67 | 1.67 |

### Mask-attention の反転によるタスク達成率の比較

Mask-attention の反転例を図 6 に示す．Mask-attention を反転した場合と，Mask-attention を反転しない場合のタスク達成率を表 3 に示す．表 3 から，Mask-attention を反転しない場合のタスク達成率より，Mask-attention を反転した場合のタスク達成率が 50pt 低下した．これは，獲得した Mask-attention が行動選択に有効な領域を表しており，ロボット動作に対する判断根拠の解析に有用であると言える．

## 5.　おわりに

本研究では，画像入力の深層強化学習モデルに Mask-attention 機構を導入したネットワークモデルを提案した．提案手法では，Mask-attention 機構を導入し，従来手法と同等の精度を獲得した．これにより，深層強化学習を用いたロボットマニピュレーションにおいて視覚的説明を実現した．今後は，オブジェクトが重なり合った状態での判断根拠の可視化と把持対象物体を変更しての実験，実機での把持実験を目指す．

## 参　考　文　献

[1] Mnih, *et al.*“Human-level control through deep reinforcement learning”, nature, pp. 529–533, 2015.
[2] R. Sutton, *et al.* “Reinforcement Learning: An Introduction”, A Bradford Book, 1998.
[3] Watkins, *et al.* “Q-learning”, Machine learning, pp. 279–292, 1992.
[4] D. Kalashnikov, *et al.*, “Scalable Deep Reinforcement Learning for Vision-Based Robotic Manipulation”, CoRL, 2018.
[5] D. Quillen, *et al.* “Deep Reinforcement Learning for Vision-Based Robotic Grasping: A Simulated Comparative Evaluation of Off-Policy Methods”, IEEE International Conference on Robotics and Automation , pp. 3651-3657, 2019.
[6] Y. Fujita, *et al.*, “Distributed Reinforcement Learning of Targeted Grasping with Active Vision for Mobile Manipulators”, IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.9712-9719, 2020.
[7] S. James, *et al.*, “Q-attention:Enabling efficient learning for vision-based robotic manipulation.” , IEEE Robotics and Automation Letters, 2022.
[8] S.Greydanus, *et al.*,“Visualizing and understanding atari agents”,Proceedings of Machine Learning Research, 2018.
[9] H. Itaya, *et al.*,“Visual Explanation using Attention Mechanism in Actor-Critic-based Deep Reinforcement Learning” , The International Joint Conference on Neural Networks, 2021.
[10] Mnih, *et al.* “Asynchronous Methods for Deep Reinforcement Learning”,Proceedings of Machine Learning Research, 2016.
[11] J. Schulman, *et al.*, “Proximal policy optimization algorithms” arXiv preprint , arXiv:1707.06347, 2017.
[12] R. Araki, *et al.*,“ARC2017 RGB-D Dataset for Object Detection and Segmentation”, Late Breaking Results Poster on International Conference on Robotics and Automation, 2018.
[13] V. Makoviychuk, *et al.*, “Isaac Gym: High Performance GPU Based Physics Simulation For Robot Learning”, Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.