アテンションマップを用いた深層強化学習による相手に合わせた動作の調整

EP18018 岡田透依　　指導教授：山下隆義

1. はじめに

深層強化学習は，高い報酬が得られる行動を選択する方策を学習する．深層強化学習で学習したエージェントは，将棋や囲碁などのゲーム環境において人間を超える性能を発揮している．一方，深層強化学習を対戦型のゲームに適用した時，報酬を最大化する行動を選択する為，相手に合わせた動作を行うことができない．そこで，本研究では深層強化学習で相手の能力に合わせて動作を選択する手法を提案する．提案手法は，行動の判断根拠となるアテンションマップに着目し，アテンションマップを調整することで，相手の能力に合わせた動作の調整を行う．

2. Mask-Attention A3C(Mask A3C)

Mask A3C[1] は，A3C に Attention 機構を導入した深層強化学習法である．Attention 機構を導入することで，行動の判断根拠の可視化が可能となる．行動の判断根拠を可視化することで，学習したネットワークの信頼性や，選択した行動が間違っていた場合の原因究明に役立つ．Mask A3C のネットワーク構造は，Feature extractor，Value branch，Policy branch，Attention 機構から構成される．Mask-Attention は Feature extractor の特徴マップに対して，1 × 1 ×チャンネル数の畳み込み層と sigmoid 関数を介して求める．Attention 機構は，Policy branch 内における中間層の特徴マップに対し，Mask-Attention を用いたマスク処理を行う．

3. 提案手法

Mask A3C は，Mask-Attention を特徴マップにマスク処理し，Policy branch に与える．そのため，Mask-Attention の注視領域を変えることで，選択される行動が変化する．本研究ではこの特性を利用して，1. 相手とのスコア差，2. 状態価値の二つに着目して Mask-Attention を調整して相手に合わせた行動を行う手法を提案する．

3.1. 相手とのスコア差による調整

相手とのスコアを拮抗させるように動作させるために，直接スコア差に着目する．スコア差に対する閾値を表 1 に示す．スコア差が優勢であれば弱く，劣勢であれば強く，拮抗状態であれば相手と同等の強さになるように Mask-Attention を調整する．

表 1：スコア差に対する閾値

スコア差	閾値
3 点↑	0.995
拮抗	0.9
3 点↓	調整無し

3.2. 状態価値による調整

状態価値は今画面のどこが得点につながるかを表している．そこで，得点を取らないような状態価値にすることで，相手と長く試合を続けられるように調整する．状態価値に対する閾値を表 2 に示す．状態価値が高ければ弱く，低ければ強く，その中間であれば相手と同等の強さになるように Mask-Attention を調整する．

表 2：状態価値に対する閾値

状態価値	閾値
1.3 ↑	0.995
1.3 ～ 0.8	0.9
0.8 ↓	調整無し

3.3. 提案手法の流れ

提案手法の流れを以下に示す．

step1:Mask A3C を学習し，ベースモデルを獲得

step2: 調整に使用する閾値のサンプリング

Mask-Attention の調整に使用する閾値ごとのモデルの強弱を確認し，強さに合った閾値をサンプリング

step3: 状況に応じて Mask-Attention を調整

スコア差や状態価値の状況に応じて閾値を決定．Mask-Attention の値が閾値以上の場合，値を 1.0 に，閾値未満の場合，値を 0.001 に調整して推論

4. 評価実験

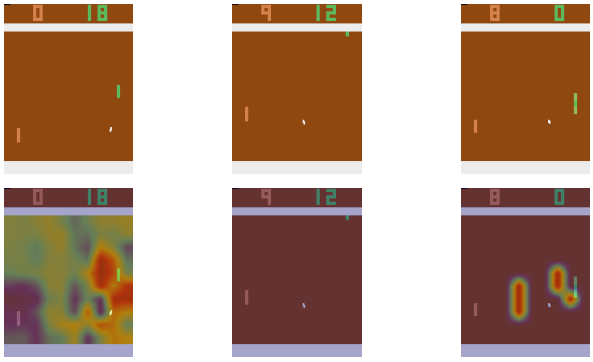
提案手法により，相手の能力に合わせた動作の調整が可能であるかの評価を行う．

4.1. 実験概要

Atari 環境における Pong を対象とし，学習を 8000 エピソード行ったモデルを使用する．Pong はボールをパドルで打ち合い，どちらかが 21 点先取で勝利となるゲームである．Mask-Attention の調整に使用する閾値の決定方法として，相手とのスコア差と状態価値の 2 通りを比較する．比較実験では，それぞれ 10 エピソード行い，最終的なスコア差と総 step 数，1 ゲームにかける step 数の平均を出力する．最終的なスコア差により相手とどのくらい点数が拮抗しているか，step 数により相手とどのくらい長くゲームが続けられたかの 2 つの観点で評価を行う．

4.2. 実験結果

通常の Mask A3C と，調整を加えた 2 つの Mask A3C の各 1500step 時における Mask-Attention を重ねたゲーム画面を図 1 に示す．図 1 より，Mask-Attention の調整が行えていることが確認できる．



(a) 調整無し　(b) スコア差　(c) 状態価値

図 1：Mask-Attention(1500step 時)

得られた最終的なスコア差と総 step 数，1 ゲームにかける step 数をまとめた結果を表 3 に示す．表 3 より，スコア差の状況に合わせて閾値を変化させた場合，最終的なスコア差が 0 に近いものとなったが，1 ゲームにかける step 数が他手法と比較し，減少していることが確認できる．これはスコア差だけの拮抗に注力し，ラリーを行わないためであると考えられる．状態価値の状況に合わせて閾値を変化させた場合，スコア差が 0 に近いとは言えないが，1 ゲームにかける step 数が他手法と比較し，増加していることが確認できる．図 1 より，1500step 時におけるゲーム数も 8 ゲームと，他手法よりも少ないことが確認できる．これは 1 点を取る間をより長くすることに注力しているためであると考えられる．スコア差で閾値を変化させることで，相手との点数を拮抗することが可能となり，状態価値で閾値を変化させることで，より長く試合を続けることが可能となった．

表 3：実験結果

	スコア差	総 step	step/1 ゲーム
調整無し	21	1827	87.03
スコア差	0.7	2588.9	66.21
状態価値	-6.3	2738.3	101.80

5. おわりに

本研究では，Mask-Attention による相手の能力に合わせた動作の調整手法を提案し，スコア差状況によりスコア差の拮抗，状態価値の状況により長く試合を続けることに成功した．今後の課題としては，これらの両立が挙げられる．

参考文献

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