Attention機構を導入したA3Cの提案

○福井宏　平川翼　山下隆義　藤吉弘亘 (中部大学)

1. はじめに

深層学習をベースにした強化学習法は，ビデオゲームやロボットにおける自律動作の獲得が可能となり，高い性能を達成している [1, 2, 3]．深層強化学習は，環境とその環境下における動作から報酬を算出してネットワークを最適化しており，エージェントが高い報酬を得られるように行動を学習する．Asynchronous Advantage Actor-Critic (A3C) [4] は，非同期学習，Advantage による学習，Actor-Critic 法の 3 要素を強化学習に取り入れたアプローチである．学習におけるこれらの 3 要素は，非同期な複数環境の効率的な学習，長期的な報酬を考慮した学習，推定した報酬のばらつきに頑健な行動の学習が可能であり，様々な自律動作を獲得するタスクに応用されている．しかし，強化学習ではどのような判断で自律動作を実現しているのかが不明であるという問題がある．

一方，深層学習による認識手法では，ネットワークがどのような意図で認識結果を出力したのかを解析する研究が取り組まれている．画像認識では，Global Average Pooling (GAP) [5] から得られる特徴マップの応答値を用いて特定カテゴリの注視領域を出力する Class Activation Mapping (CAM) [6, 7] 等が提案されている．CAM の特定物体における注視領域を応用した手法として Attention Branch Network (ABN) [8] がある．ABN は，GAP から得られる特徴マップを Attention map として使用し，Attention 機構へ応用することで物体認識性能を向上している．Attention map は，入力データからどの部分を注視したかを示す領域である．ここで，ABN はネットワークの上位層が Attention map を出力する Attention branch と，出力された Attention map から最終的な認識結果を出力する Perception branch により構成されている．

本研究では，自律動作の獲得の性能向上，及び推論時の動作における注視領域を視覚化するために，ABN の Attention 機構を導入した A3C を提案する．A3C は，行動と状態価値を出力するために Multi-task Learning を導入しており，出力される状態価値は入力の状態がどの程度報酬に貢献するのかを数値的に示している．そのため，状態価値を ABN と同様に Attention map を GAP により出力することで，特定の環境から状態価値の高い局所的な領域を注視した Attention map を得ることができる．提案手法である，ABN の Attention 機構を導入した A3C を図 1(a) に示す．図 1(a) のように A3C の上位層を ABN のようにブランチを構築することで，図 1(b) のような報酬の高い領域を注視しながら行動を End-to-End に学習，獲得できる．評価実験では，Open AI gym [9] のゲーム環境を使用し，提案手法の有効性を検証する．

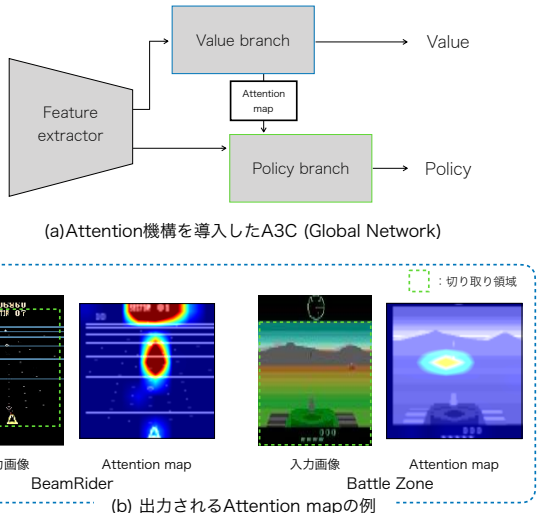

図 1　提案手法のネットワーク構造と Attention map

2. A3C

A3C は，複数環境の非同期な学習 (Asynchronous)，数ステップ先の報酬を考慮した学習 (Advantage)，エージェントの行動と状態価値の出力による安定した学習 (Actor-Clitic 法) の 3 つ要素を導入した強化学習法である．

Asynchronous　A3C のネットワーク構造は，大元の Global Network と，異なる環境を複数のエージェントで学習する worker から構成される．worker は，Global Network のコピーであり，与えられた環境を入力として学習される．それぞれの異なる環境で学習された worker のパラメータは，Gobal Network のパラメータに非同期で反映される．非同期な学習は異なる環境での経験を Global Networ へに取り入れれるため，効率的かつ高速に学習できる．

Advantage　Advantage では，式 (1) のように数ステップ先の報酬 r を考慮してネットワークを学習する．ここで，γ は時間割り率であり，t ステップ後の報酬をどの程度反映させるかを決定する係数である．ここで，$\mathbf{s}$ と $\mathbf{a}$ はそれぞれ状態と行動を示している．一般的な Deep Q-Network は 1 ステップ先の報酬のみ考慮するため学習が不安定であったが，数ステップ先の報酬を考慮することで，より安定した学習を可能にしている．

$$Q(\mathbf{s}_t, \mathbf{a}_t) \rightarrow r_t + \gamma r_{t+1} + \gamma^2 \max Q(\mathbf{s}_{t+2}, \mathbf{a}) \quad (1)$$

Actor-Clitic 法　Actor-Clitic 法は，行動 (Actor) と状態価値 (Critic) を出力とし，ネットワークを学習する強化学習法の一つである．Deep Q-Network では出力された行動から報酬を求めるため，報酬のばらつきが行動の学習に悪影響を及ぼしていた．しかし，行動と状態価値を同時に学習することで，報酬の変化が行

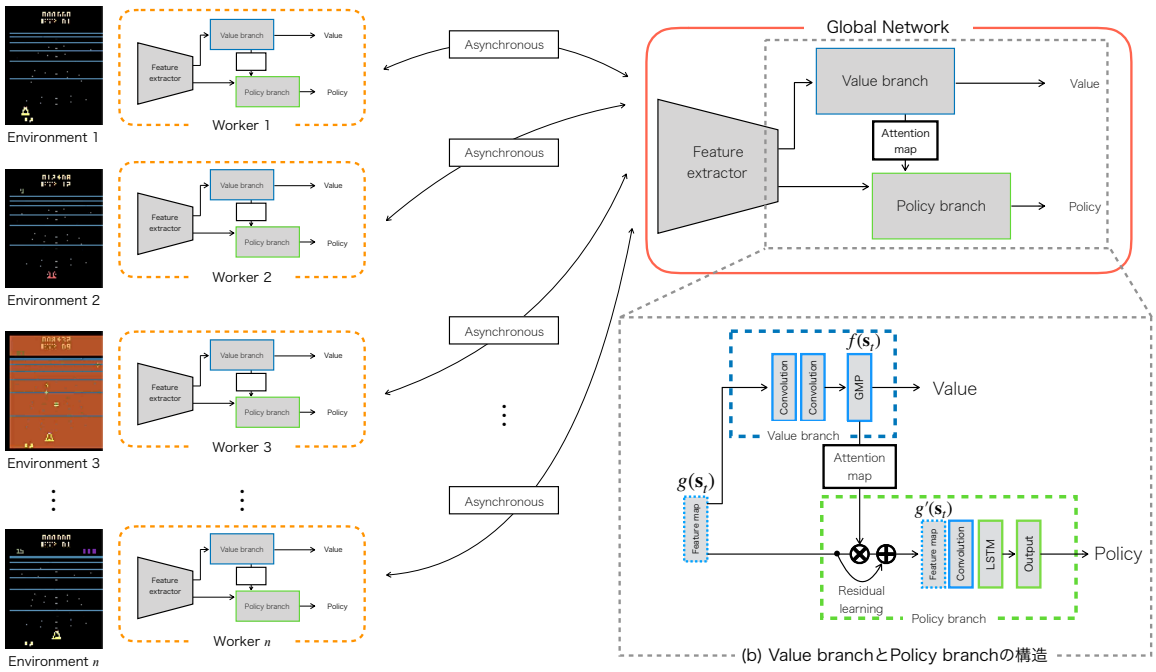

図 2　提案手法の構造

動に悪影響を与えにくくしている．

3. 提案手法

3.1 Attention 機構を導入した A3C

ABN ベースの Attention 機構を導入した A3C の構造を図 2 に示す．図 2 のように提案手法の学習アルゴリズムは従来の A3C と同様であるが，ネットワークの上位層が ABN のように 2 つのブランチで構築されている．ABN は上位層に Attention map を出力する Attention branch と，最終的な出力結果を出力する Perception branch が配置されており，両ブランチを同一のクラスラベルで学習する．Attention 機構を導入した A3C では状態価値を推定する過程で Attention map を出力し，出力された Attention map を考慮して行動を出力する．そのため，本論文では状態価値と Attention map を出力するブランチを Value branch，Attention map を考慮して行動を出力するブランチを Policy branch と呼称する．また，A3C は Long-Short Term Memory (LSTM) [10] を併用することで性能を大幅に向上できるため，提案手法においても LSTM を導入する．

本手法の流れは，はじめに Global Network を構築した後に，Global Network のコピーである worker を複数構築する．このとき，Feature extractor，Value branch，Policy branch を用いてネットワークを構築する．そして，各 worker に対して異なる環境がそれぞれ入力される．入力される環境から，Feature extractor により特徴マップを抽出する．特徴マップは，はじめに Value branch へ入力されて Attention map と状態価値を出力する．そして，出力された Attention map と Feature extractor から出力された特徴マップは，Policy branch に入力される．Policy branch では，Attention map を特徴マップに反映させ，行動を出力する．これにより，Policy branch では Attention map を用いることで価値の高い領域を注視しながらが行動を学習，獲得できる．本章では，提案手法である Attention 機構を導入した A3C の詳細な処理について述べる．

3.2 Value branch

Value branch と Policy branch の構造を図 2(b) に示す．Value branch の畳み込み層は，全て 3×3 のカーネルを複数枚有する畳み込み層で構成されている．これらの畳み込み層のチャンネル数は，Policy branch の全結合層のユニット数と同様であり，それぞれ 64 と 512 チャンネルである．これらの処理は Fully Convolutional Network (FCN) [11] で用いられている方法であり，畳み込み層をベースに全結合層を構築することで Feature extractor の特徴マップを縮小せずに Attention map を生成できる．

A3C は出力する行動と状態価値の 2 つのタスクを Multi-task Learning で学習することで，2 つのタスクを互助しながら学習する．しかし，提案手法の A3C は Feature extractor の後層が状態価値を出力するブランチと行動を出力するブランチに分離されるため，両タスク間における互助を打ち消してしまい，学習が困難になる．この問題を解決するために，Value branch から状態価値を出力する際に，GAP ではなく式 (3) の Global Max Pooling (GMP) を用いる．GAP では Attention map の平均を状態価値として出力するため，間接的に出力する状態価値を Policy branch へ反映することになる．しかし，GMP では Attention map の最大値を状態価値とするため，環境 $\mathbf{s}_t$ における状態価値を直接的に Policy branch に与える事ができる．

$$Value_t = \max f(\mathbf{s}_t) \tag{2}$$

3.3 Policy branch

図 2(b) のように Value branch から出力された Attention map は，Feature extractor から得られた特徴

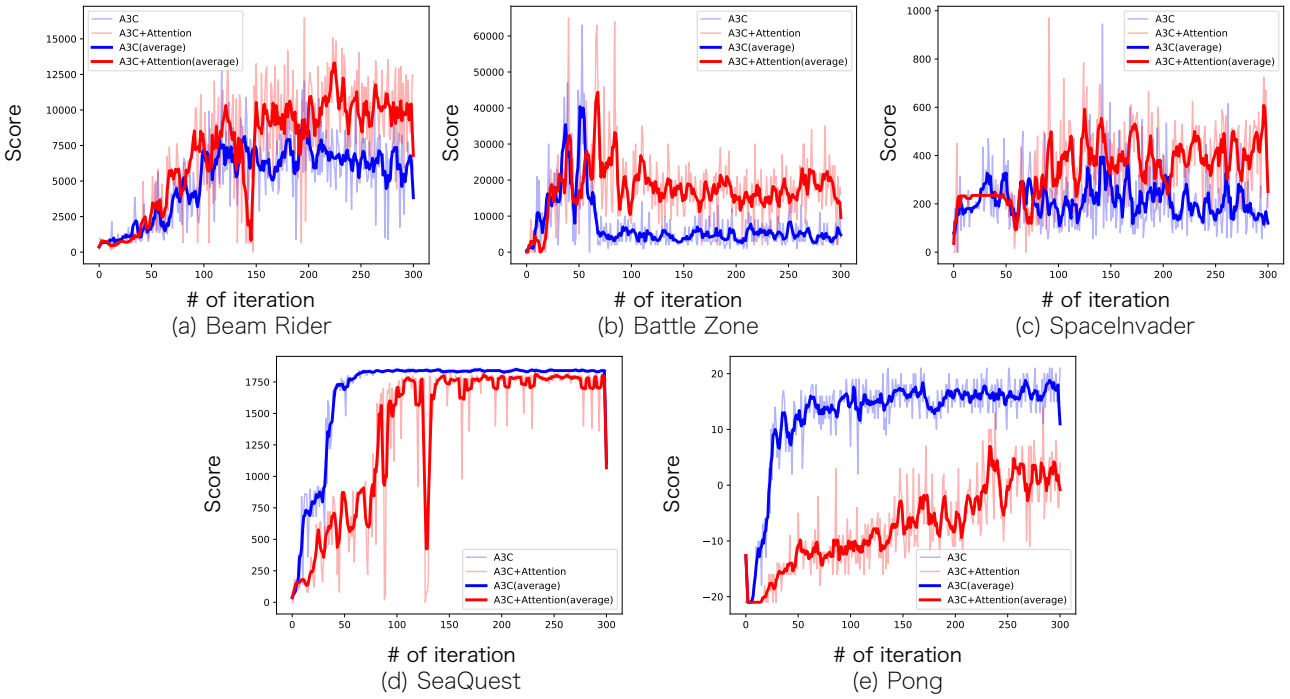

図 3　Open AI gym における各ゲームのスコアの推移

マップ $g(\mathbf{s}_t)$ に反映され，新たな特徴マップ $g'(\mathbf{s}_t)$ が出力される．Attention map の反映には，式 (3) のような Residual 機構 [12] を導入する．この構造は Residual Attention Learning と呼ばれており，Attention map による特徴マップの消失を抑制できる [13]．そして，Attention map を施した特徴マップを畳み込み層及び LSTM に入力することで，行動を出力する．

$$g'(\mathbf{s}_t) = (1 + f(\mathbf{s}_t)) * g(\mathbf{s}_t) \tag{3}$$

4.　評価実験

提案手法である Attention 機構を導入した A3C の有効性を確認するために，Open AI gym [9] のゲーム環境を用いて評価実験を行う．今回の実験で検証するゲームは，“Beam Rider”，“Battle Zone”，“Space Invader”，“Sea Quest”，“Pong”である．これらのゲームでは操作画像をネットワークに入力し，ゲーム内のキャラクタや乗り物をエージェントとして操作することでゲームのスコアを獲得する．入力する画像は任意の領域を切り出した後に，80×80 のグレースケール画像に変換する．

使用する A3C は，畳み込み層 4 層と LSTM，全結合層から構成されている．Attention 機構を導入した A3C は，従来の A3C をベースに構築する．はじめに，Feature extractor を畳み込み層 2 層で設計し，Value branch を畳み込み層 3 層と GMP，Policy branch を畳み込み層 1 層と LSTM，出力層で設計する．学習時のパラメータは，学習係数を 0.0001，時間割引き率 γ を 0.99，worker の数を 32 とする．学習の終了条件は，Global Network のエピソード数を 300 回，worker の更新回数を 500 回とする．

4.1　性能比較

図 3 に Open AI gym の環境において得られた従来の A3C 及び提案手法の報酬の推移を示す．横軸は Global Network のエピソード数を示しており，縦軸は獲得したゲームのスコアを示している．ここで，グラフ中の半透明の実線が各エピソードで得られたスコア，濃い実線は得られたスコアの移動平均を示している．

図 3 の結果から，(a)BeamRider，(b)BattleZone，(c)SpaceInvader においては提案手法の Attention 機構を導入した A3C (A3C+Attention) が従来の A3C より高いスコアを獲得できている．一方で，(d)Sea Quest と (e)Pong は従来の A3C より獲得できたスコアが低い事が確認できる．

表 1 に，評価時における最終的な各ゲームのスコアを示す．評価時は，学習したそれぞれの A3C に対して最大エピソード数を 10,000 に設定して 100 回評価する．そして，獲得したスコアの最大と平均を用いて従来法と提案手法の性能を比較する．表 1 より，最終的なスコアの比較においても図 3 と同様の傾向であり，BeamRider，BattleZone，SpaceInvader では従来の A3C より高いスコアを獲得できた．Sea Quest では，従来の A3C とほぼ同等のスコアを獲得できた．Pong においては，従来の A3C よりスコアが大幅に低下していることが確認できる．

4.2　Attention map の可視化

提案手法により得られた Attention map を可視化する．各ゲームの Attention map の可視化結果を図 4 に示す．(a)Beam Rider や (b)Battle Zone，(c)SpaceInvader，(d)Sea Quest では，敵に対して Attention map が強く反応している．特に，(a)Beam Rider と (d)Sea Quest では他のゲームとは異なる傾向が見られた．(a)Beam Rider では，画面左上に出現

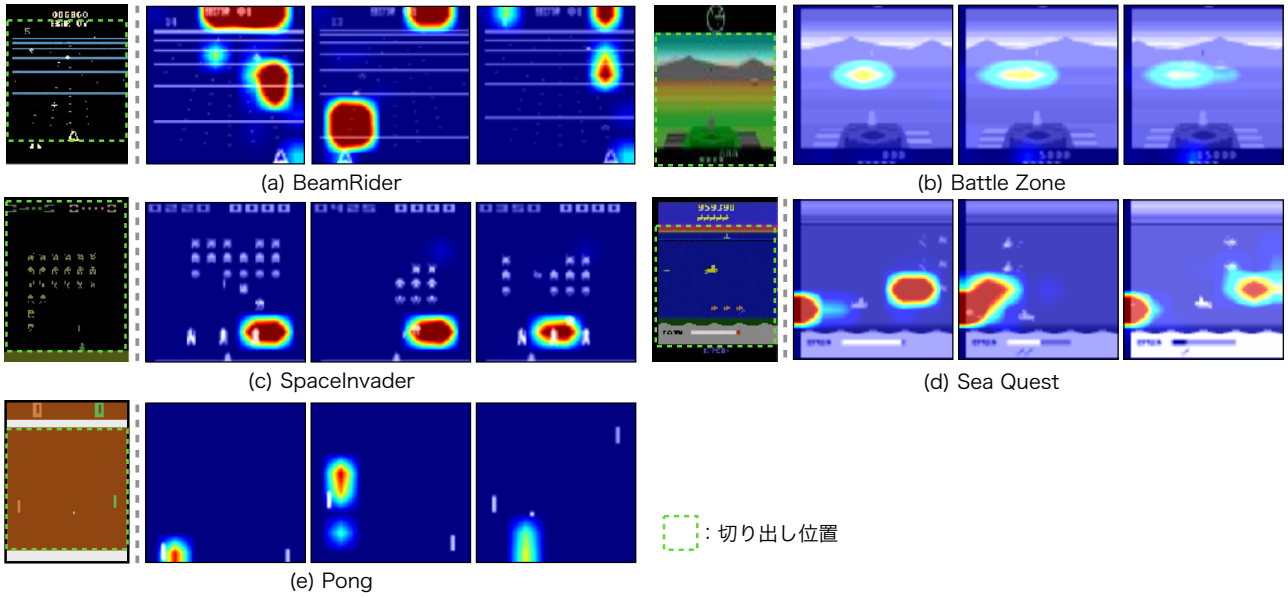

図 4　Attention 機構を導入した A3C の Attention map の例

表 1　評価時におけるゲームスコアの比較

	最大スコア		平均スコア	
手法	A3C	A3C+Att.	A3C	A3C+Att.
Beam Rider	12,940	**15,078**	7,664.0	**10,071.1**
Battle Zone	80,000	**89,000**	37,820.0	**48,690.0**
SpaceInvader	665	**925**	239.8	**412.8**
Sea Quest	**1,860**	1,820	**1840.2**	1758.8
Pong	**21**	14	**16.9**	4.6

する特定の数字に対しても Attention map が強く反応した．この数字は，スコアの高い敵が出現するまでのカウントダウンであり，この数字が小さくなるに連れて Attention map が強く反応した．(d)Sea Quest では，画面右側から敵が発生しやすい．そのため，自機は右側を向いている場合が多く，図 4(d) の画面左下の Attention map が強く反応している領域に敵機が出現した際に，自機が画面左側に移動して敵を追撃している．

(e)Pong では主にボールに対して Attention map が反応しており，特に相手から点を獲得した際に Attention map が強く反応する．しかし，Pong は他のゲームとは異なり，行動を起こしてしばらく時間を置いた後にスコアが加点されるか否かが決定する．そのため，スコアが加算されて Attention map が強く反応した場合でも，エージェントの行動に対しては全く影響が及ばない．この原因により，従来の A3C より性能が大幅に低下したと考えられる．

5.　おわりに

本研究では，Attention 機構を導入した A3C を提案した．提案手法では，ABN の Attention 機構を A3C に応用することで，特定のゲームの性能を向上し，エージェントの制御に関する解析を行った．しかし，得られる報酬と出力する行動のタイミングの差が原因で，一部のゲームの性能が大幅に低下した．そのため，今後は出力される状態価値に影響されない Attention map を出力する Value branch の提案を行う．

参　考　文　献

[1] V. Mnih *et al.*, “Playing atari with deep reinforcement learning,” *arXiv preprint arXiv:1312.5602*, 2013.
[2] D. Silver *et al.*, “Mastering the game of go without human knowledge,” *Nature*, vol. 550, pp. 354–, 2017.
[3] Y. F. Chen *et al.*, “Socially aware motion planning with deep reinforcement learning,” *International Conference on Intelligent Robots and Systems*, 2017.
[4] V. Mnih *et al.*, “Asynchronous methods for deep reinforcement learning,” in *International Conference on Machine Learning*, 2016.
[5] L. Min, C. Qiang, Y. Shuicheng, “Network in network,” *International Conference on Learning Representations*, 2014.
[6] B. Zhou *et al.*, “Learning deep features for discriminative localization,” *Computer Vision and Pattern Recognition*, 2016.
[7] S. Ramprasaath, R. *et al.*, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *International Conference on Computer Vision*, 2017.
[8] 福井宏 *et al.*, “Global average pooling の特性を用いた attention branch network,” 2018.
[9] G. Brockman *et al.*, “Openai gym,” 2016.
[10] S. Hochreiter, J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[11] J. Long, E. Shelhamer, T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Computer Vision and Pattern Recognition*, 2015.
[12] K. He *et al.*, “Deep residual learning for image recognition,” *Computer Vision and Pattern Recognition*, pp. 770–778, 2016.
[13] F. Wang *et al.*, “Residual attention network for image classification,” in *Computer Vision and Pattern Recognition*, 2017.