5M-09

転移学習による Deep Q-Network の学習高速化に向けた検討

足立 一樹† 佐々木 勇人‡ 中田 雅也§ 濱津 文哉§ 濱上 知樹§
†横浜国立大学 理工学部 ‡横浜国立大学 大学院工学府 §横浜国立大学 大学院工学研究院

1 まえがき

Q-Learning[1] に深層学習を取り入れた手法である Deep Q-Network（DQN）[2] には，従来の Q-Learning では扱いきれない，画像のような高次元の観測を直接扱うことができるという利点がある．一方，課題として学習を行う際には膨大な回数のエピソードを繰り返す必要がある．この課題に対処するために，別のタスクで学習済みの畳み込みニューラルネットワーク（Convolutional Neural Network: CNN）[3] を利用した転移学習が提案されている [4]．転移学習によりエージェントはタスクに有用な特徴抽出が可能な状態から学習を開始できると考えられる．しかし，どのような場合に転移が有効であるか，CNN のどの層までの特徴は転移に有用であるかは明らかになっていない．本稿では転移を行う CNN の層数を変化させることで学習回数や得られる報酬にどのような影響が現れ，転移された特徴がどのように利用されるのかを実験により調査する．

2 DQN の転移学習に関する従来研究

Yunshu Du らによる従来研究 [4] では Atari ゲーム（Breakout, Pong: 図 1）をタスクとして DQN の転移学習が行われている．

DQN で Atari ゲームの学習に用いられる CNN は，図 2 のようにゲーム画面を状態 s として入力し，プレイヤーの各行動 $a_1, \ldots, a_{|\mathcal{A}|} \in \mathcal{A}$ に対する Q 値 $Q(s, a_1; \theta), \ldots, Q(s, a_{|\mathcal{A}|}; \theta)$ がそれぞれ出力される構造になっている．従来研究ではこの CNN において，3 層ある畳み込み層のパラメータを一方のタスクで学習済みのものからコピーし，それを他方の初期値として，類似した要素を持つもう一方のタスクの学習が行われている．これにより，エージェントがタスクに有用な事前知識を持った（特徴抽出が可能な）状態から学習を開始できると考えられている．しかし，具体的にどのような場合に転移が有効であるか，また転移学習を行うことによって CNN のパラメータがどのように変化するかなどについては明らかではないため，これらを更に深めていく必要がある．

3 実験

DQN におけるパラメータの転移学習の有効性を確認するとともに，転移された畳み込み層で行われる特徴抽出がどのように利用されているかを調査するために以下の実験を行った．

Breakout, Pong のタスクの学習をそれぞれ行った後，それらを学習した CNN を用いて [4] と同様に転移学習

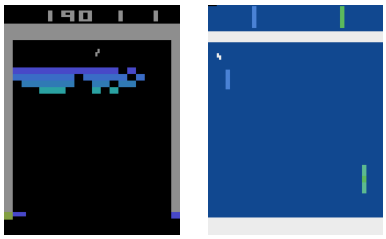

図 1: Atari ゲーム

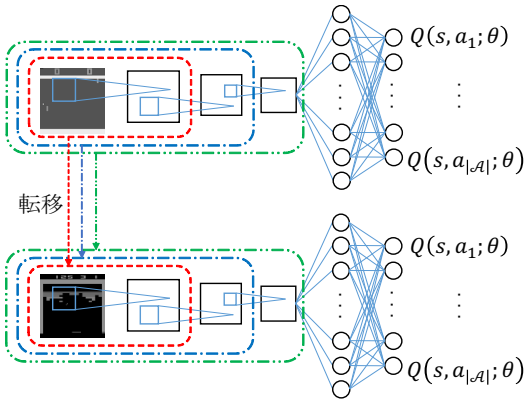

図 2: Atari ゲームの学習に用いられる CNN と転移学習の概要

の実験を行った．これらのタスクには「バーを動かしてボールを打ち返す」という類似点があるため転移の有効性を確認するのに適切であると考えられる．

図 2 のように，Pong を学習した CNN の畳み込み層のパラメータをコピーし，それを初期値として Breakout の学習を行った．これにより，タスクに有用な特徴抽出を行える状態から学習を開始できると考えられる．学習開始時にコピーする畳み込み層を 3 層あるうちの 1 層目のみ，1,2 層目，1,2,3 層目と変化させて学習の様子の変化を観察した．

また，転移元と転移先のタスクを逆にした実験も同様に行った．

ハイパーパラメータの設定は文献 [2] と同様とし，CNN の学習はミニバッチ学習で行った．バッチサイズは 32 で，最適化アルゴリズムには RMSpropGraves を用い，割引率は $\gamma = 0.99$ とした．

4 実験結果と考察

4.1 転移学習

Pong で学習を行ったパラメータを Breakout の学習に転移（Pong → Breakout）した結果を図 3 に，Breakout で

A learning acceleration of Deep Q-Network with Transfer Learning
†Kazuki ADACHI ‡Hayato SASAKI §Masaya NAKATA §Fumiya HAMATSU §Tomoki HAMAGAMI
†Department of Science and Engineering, Yokohama National University
‡Graduate School of Engineering, Yokohama National University
§Faculty of Engineering, Yokohama National University

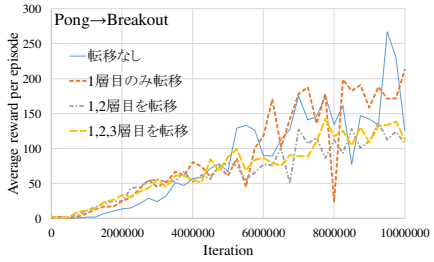

図 3: 転移学習（Pong → Breakout）の結果

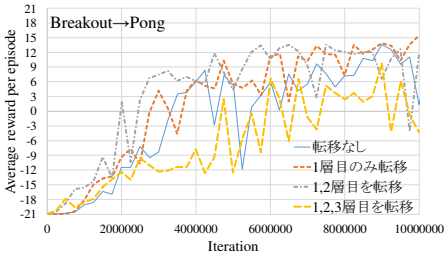

図 4: 転移学習（Breakout → Pong）の結果

学習を行ったパラメータを Pong の学習に転移（Breakout → Pong）した結果を図 4 にそれぞれ示す．

ただし，「転移なし」は転移を行わずランダムな初期パラメータから学習を行ったものを表す．

図 3 の結果では，学習初期は大きな変化はなかったが，転移を行った場合の方が早く高得点を得られるようになった．しかし，学習後期では最終的な性能は転移なしの場合と同程度であった．

また，図 4 の結果においても学習初期は転移を行った場合のほうが早く高得点を得られるようになった．学習後期では 3 層すべて転移したものは性能が下がったが，他の場合は転移なしの場合と同程度の性能となった．

4.2 パラメータ変化量の推移

転移を行った畳み込み層がどの程度タスクに有用であったかを評価するために，畳み込み層のパラメータ（重み W, バイアス $\boldsymbol{b}$）の転移直後のパラメータ $W_0, \boldsymbol{b}_0$ に対する変化量 D を (1) 式で計算し，イテレーション回数に対して層ごとにプロットした．

$$D = \frac{\|W - W_0\|_{\mathrm{F}}^2}{\text{input} \times \text{output} \times \text{ksize}^2} + \frac{\|\boldsymbol{b} - \boldsymbol{b}_0\|^2}{\text{output}} \qquad (1)$$

汎用性の高い特徴抽出が行われている層であれば転移直後のパラメータからの変化は小さいと考えられるため，転移した層で抽出される特徴がどの程度一般的か評価することができる．

Pong → Breakout と Breakout → Pong で，それぞれ 1,2,3 層目の転移を行った場合の D の推移を図 5,6 に示す．

Pong → Breakout と Breakout → Pong のいずれの転移においても，最も汎用性の高いと思われた 1 層目のパラメータの変化が最も大きく，上位層のパラメータの変化は小さかった．これは「ボール」,「バー」などのタスクに特化した特徴であっても抽出した特徴が利用しやすかったためであると考えられる．

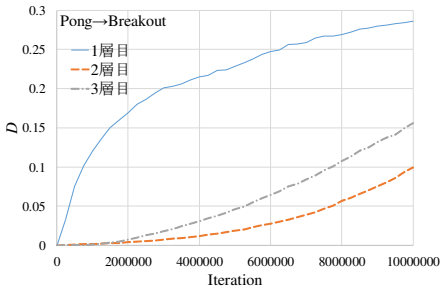

図 5: Pong → Breakout におけるパラメータ変化量の推移

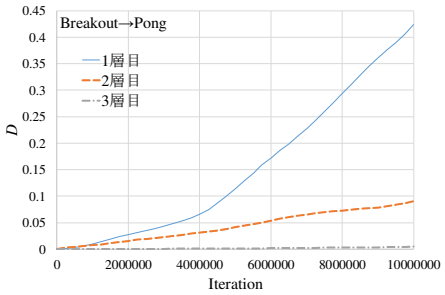

図 6: Breakout → Pong におけるパラメータ変化量の推移

5 まとめ

転移学習において畳み込み層のパラメータの変化から，転移した特徴がどの程度利用できているかを分析した．共通点を持つタスクであれば，タスクに特化した特徴抽出が行われる上位層の転移も有効であると考えられる．

今後はこれらの結果をもとに，どのようなタスクがどの程度転移に有効であるかなどを定量的に評価できる方法や，学習するタスクに有用な特徴の選択も含めて学習できるような方法などを検討していく．

参考文献

[1] Watkins, Christopher John Cornish Hellaby. “Learning from delayed rewards.” Diss. University of Cambridge, 1989.

[2] Mnih, Volodymyr, et al. “Human-level control through deep reinforcement learning.” Nature 518.7540 (2015): 529-533.

[3] LeCun, Yann, et al. “Gradient-based learning applied to document recognition.” Proceedings of the IEEE 86.11 (1998): 2278-2324.

[4] Yunshu Du, Gabriel de La Cruz, James Irwin and Matthew Taylor. “Initial Progress in Transfer for Deep Reinforcement Learning Algorithms.” Proceedings of the Deep Reinforcement Learning: Frontiers and Challenges (DeepRL) workshop at the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016)