

深層強化学習における状態系列表現の獲得に向けて

野口 直人^{*1} 甲野 佑^{*2} 高橋 達二^{*1}

^{*1}東京電機大学理工学部 ^{*2}東京電機大学大学院

1. はじめに

近年、深層学習と呼ばれる手法を用いて、画像認識の分野では、一般物体認識コンテストにおいて既存手法よりも大幅に高い成績を示し、音声認識の分野においても、高精度な学習で実用化もされている [Lecun 15]. さらに Deep Q-network (DQN) と呼ばれるゲーム画面を視覚情報として与え、試行錯誤のみから良いプレイを学習する手法も考案されており、過半数のゲームで人間の熟練者よりも高い成績を収める事が示されている [Mnih 15]. DQN では一度に認識できるゲーム画面の時系列数 (フレーム数) を固定して学習をさせるが、最適なフレーム数はゲームによって異なり、またフレーム数の増加に比例して入力ユニット数が増加するため学習中に動的に変更する事は出来ない. 深層学習では可変長の時系列データを扱うために再帰的ニューラルネットワーク (RNN) を DQN に組み込む試みはなされている [Hausknecht 15]. しかし DQN と RNN の両者とも繊細なパラメータチューニングが必要であるため、付随する課題が山積している. それに対して我々は DQN のフィードフォワードなデータの流れの中ではなく、ある一種の迂回路において時系列データとして扱いやすい表現を獲得する事が重要になるのではないかと我々は考えた. 本研究はそのような時系列データとして扱いやすい、節約的な表現 (記号化された表現) を獲得する事を目的とした Deep Q-Network with Symbolization (SDQN) を提案する.

2. Deep Q-Network

Deep Q-Network (DQN) とは画像認識に多く用いられる畳み込みニューラルネットワーク (CNN) と強化学習の行動価値近似器 (Q-network) を組み合わせた学習アーキテクチャであり、ゲームのルールを教えていない場合でも視覚 (ただし報酬であるゲームの獲得した得点の差分の認識は視覚情報ではなく直接的に与えられる) と試行錯誤のみで、高い得点を獲得でき

るゲームのプレイ方策を学習していく. 本研究で用いる DQN は入力情報としてグレースケール化した時刻 t と過去 2 フレーム ($t-1, t-2$) のゲーム画面を入力情報として受け取る (オリジナルの研究ではこれより一つ多く、合計 4 フレームである). 入力された情報から CNN 部が状態認識を行い、Q-network 部によってその状態での Q 値を算出して行動を決定する.

3. 提案手法 -SDQN-

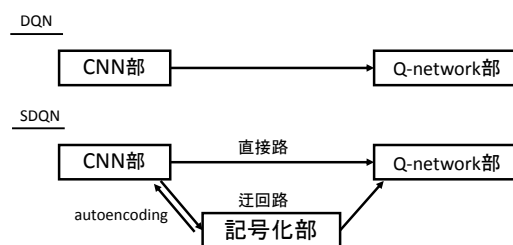


図 1: DQN と SDQN の概略図

入力の特徴をより節約して表現するような学習を行える autoencoder は、3 層ニューラルネットワークになっており、入力情報がそのまま教師信号として用いて元の入力を復元できるような、元のユニット数より圧縮された中間層の重みパラメータを学習する. そこで我々は autoencoder を DQN に組み込む事で時系列データを処理しやすい表現を獲得できると考え、DQN の CNN 部 → Q-network 部という直接路とは別に、CNN 部 → 記号化部 → Q-network 部という迂回路を設けた記号化を伴う DQN, Deep Q-Network with Symbolization (SDQN) を考案した (図 1). SDQN では迂回路の CNN 部 → 記号化部は活性化関数にシグモイド関数、誤差関数にシグモイドクロスエントロピー関数を用いて学習し、誤差逆伝播は行わないように設定している.

4. Atari 2600 実験

本研究で考案した SDQN が行動の学習を行えるか調べるため、DQN と同じく Atari 2600 のゲームを用いた実験を行った. 学習課題には Breakout (ブロック崩し) を用いた. DQN 及び SDQN のネットワークはミニバッチ学習で学習される. ミニバッチのバッチ数は 32 で、割引率 $\gamma = 0.95$, 学習率は AdaDelta によって調整される. 学習はプレイ結果の保存場所 replay memory にサンプル (s, a, r, s') が 100 サンプルた

Toward Acquisition of State Sequence Representation in Deep Reinforcement Learning.

Naoto Noguchi, Tatsuji Takahashi, School of Science and Technology, Tokyo Denki University.

Yu Kohno, Graduate School of Tokyo Denki University.

まってから開始される．また，replay memory の最大容量は 500,000 サンプルである．行動は ϵ -greedy 法によって決定され，ランダム選択確率 ϵ は ϵ 減衰法を用い，学習時間を経るごとに減少していく．減衰の速度は DQN では学習回数 0 の $\epsilon = 1.0$ で，学習回数 1,000,000 で $\epsilon = 0.1$ ，SDQN ではユニット数がほぼ 2 倍のため，学習の遅延を考慮して学習回数 2,000,000 で $\epsilon = 0.1$ になるよう線形に減衰する設定した．また，それ以上時間を経ても $\epsilon = 0.1$ からは減衰しない．DQN，SDQN 共に学習回数は 1,000,000 回とした．

4.1 結果および考察

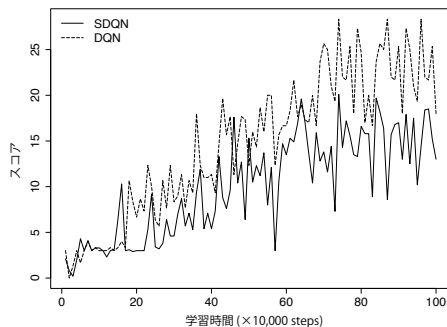


図 2: Breakout における SDQN と DQN のスコアの遷移

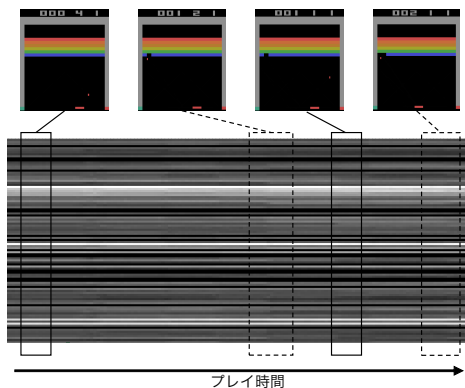


図 3: 学習回数 10000 の記号化層発火ユニット

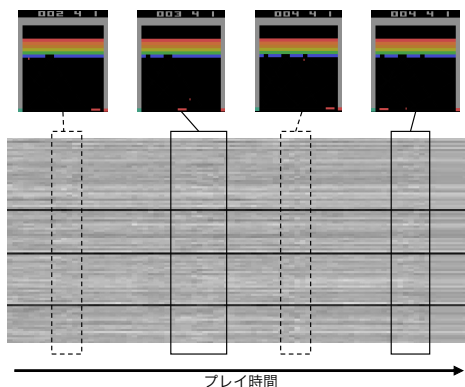


図 4: 学習回数 1000000 の記号化層発火ユニット

図 2 は Breakout での学習回数 10,000 回毎にその時点で学習されたネットワークを用いたプレイでの成

績であり，シミュレーション 10 回の結果の平均である．この結果から，やや DQN に劣るものの SDQN では学習対象となる重みパラメータ数が約 2 倍になっているのに関わらず，同等の学習速度での学習が行えていることがわかった．

さらに，記号化層のプレイ中の出力パターンを示す学習初期 (10,000 回目) の図 3 と学習後期 (1,000,000 回目) の図 4 では，明らかな反応の差が見られることが確認できた．学習初期 (図 3) ではボールとバー (プレイヤーが操作可能なボールを打ち返す部分) の位置関係，ブロックの崩れ具合に対して差がなく，プレイ時系列に対して一様な反応しか示さない．しかし学習後期 (図 4) ではボールがブロックを崩した状態 (図 4 破線範囲)，ボールがバーより右側に位置する状態 (図 4 実線範囲) それぞれに共通した特徴があり，それ以外とは異なることを確認した．

5. 結論

強化学習する深層学習の学習器の一つである DQN は入力が固定長であるために可変長の時系列データを扱えない．本研究ではその解決の前段階の処理として DQN に入力データを記号化して行う迂回路を追加した SDQN を考案した．結果として，DQN と同等の学習水準を保ちながら，記号化層での学習初期と学習後期でのユニットの出力傾向の違いや，類似した課題状況に対して，類似した出力値の変化のパターンを示すことがわかった．

現時点では，本研究で示したような課題状況の類似判断に関する可視化手法や定量的な評価法が存在しない．今後はより客観的に記号化層の能力やその改善を行うために，評価法自体について考察する必要がある．更に長期的には，RNN のように記号化層を再帰的に結合し，画面 1 フレームのみの入力であるのに，過去のフレーム全てを入力に用いて実現されるような長期的な時間感覚での学習や，画面に表れずにプレイヤーから隠れた状況を推論するような学習にも対応可能になると考えられる．

参考文献

- [Lecun 15] Y. Lecun, Y. Bengio, G. Hinton, Deep Learning, *Nature*, 521 (7553) (2015), pp. 436–444.
- [Mnih 15] V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, A. Graves, M. Riedmiller, A.K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis, Human-level control through deep reinforcement learning, *Nature*, 518 (7540) (2015), pp. 529–533.
- [Hausknecht 15] M. Hausknecht, Peter Stone, Deep Recurrent Q-Learning for Partially Observable MDPs, *CoRR abs/1511.04143* (2015).
- [LeCun 15] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature*, 521(7553) (2015), pp. 436–444.