ISSN 2186-5647

2-18

動的強化学習ネットワークにおける特徴抽出と強化学習の融合

日大生産工(院)　○小松　泰士　　日大生産工　山内　ゆかり

1. まえがき

近年，強化学習[1]を搭載した自律的にタスクを遂行するエージェントやロボットの実現が強く望まれている．その中で，環境を通じて適応的な学習が可能な強化学習が活躍している．強化学習とはエージェントの試行錯誤により与えられた報酬のみで行動を学習するアルゴリズムである．ある状態sの時にある行動aを行う時に得られる報酬の期待値をQ値として学習を行う．Q値の学習のために様々な手法が存在する．その中でも近年では深層学習と強化学習を組み合わせた深層強化学習が注目されている．その一つとしてDeepMind社によって提案されたDeep Q-Network[2]がある．Deep Q-Networkとは，学習に必要な状態空間の入力を畳み込みニューラルネットワーク(Convolutional Neural Network：CNN)[3]によって特徴抽出を行い，Q値を学習するアルゴリズムである．事前の知識情報なしで複雑な環境のルールや行動の優先度を学習することが可能となっている．Deep Q-Networkはフィルタによる畳み込みを用いて特徴抽出を行っているが，事前に既定したフィルタ層の数やフィルタのサイズに従って特徴抽出後の次元数が固定される．フィルタ層の深さによっては特徴抽出後の次元数が膨大になり処理に時間がかかる問題点が存在する．

そこでDeep Q-Networkの特徴抽出部にオンライン学習により不必要な状態空間の削除を行う自己増殖型ニューラルネットワーク(Self-Organizing Incremental Neural Network：SOINN)[4]を採用したDynamic Q-Network[5]が提案された．SOINNを用いることで，学習が進んだ際に不必要になった状態空間を自動で除去が可能なため，動的に特徴抽出後の次元数を増減させながら学習する．また，SOINNの特徴抽出部は全結合層の勾配降下法とは独立して学習することでネットワークの層の深さを抑え，学習効率の向上を目指している．

本研究では，特徴抽出部のSOINNと強化学習を行うQ-Networkの融合に向けて，2つの改良手法を提案する．まず動的に特徴抽出を行うSOINN部でノード削除時において，Q-Networkで結合荷重のコサイン類似度が最も近いノードの重みに影響を残すことで，ノード削除による学習率の低下を防ぐ．また，Q-Networkで学習により重要度が低下した，あるいは冗長となったと考えられるノードを選出し，SOINN空間のノードの削除候補に取り入れることで，計算量の削減と過学習の防止を目指す．Reversiを実験環境とし，処理時間及び勝率により提案手法の有効性を示す．

2. 従来研究

2.1. Deep Q-Network

Deep Q-Network[2]とはDeepMind社によって提案された深層学習とQ-Networkを組み合わせた学習アルゴリズムである．通常のQ-Networkとは異なり，状態の入力次元から畳み込みニューラルネット(Convolutional Neural Network：CNN)[3]による特徴の抽出の後に全結合層の学習を行う．入力の似ている状態を特徴ごとに分類分けを行うことで，事前の知識情報なしで複雑な環境のルールや行動の優先度を自動で学習することが可能となっている．図1に“Human-level control through deep reinforcement learning”の論文内で設計されていたDeep Q-Networkのモデルを示す．

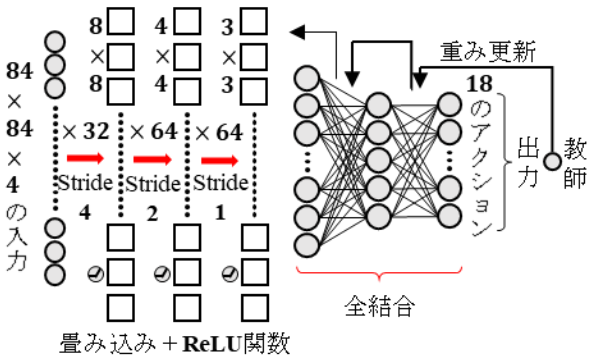

図1 Deep Q-Networkのモデル

Atari2600のゲーム画面のピクセル84×84の4フレームを入力としている．3層の畳み込み層を通して特徴抽出を行い，全結合層の入力を決定している．全結合層の出力にはAtari2600の操作方法である18アクションを出力ノード数としている．

時間tの時の状態s_tの場合の行動a_t番目の出力ノードO_{a_t}のニューラルネットワークの重みω_{ij}，及び畳み込みに用いるフィルタh_{pq}の更新には教師T_{a_t}が必要である．勾配計算で扱う教師の設

Feature Extraction by Competitive Learning for Dynamic Q-Network

Taishi KOMATSU and Yukari YAMAUCHI

定式を式(1)に示す．ここでαは学習率，γを割引率とする．

$$T_{a_t} = O_{at}(s_t) + \alpha[r_{t+1} + \gamma max_{a'} O_{a'}(s_{t+1}) - O_{a_t}(s_t)] \quad (1)$$

CNN による特徴抽出部分では$I \times J$の出力u_{ij}を求める場合，入力xを$P \times Q$のフィルタh_{pq}を通してストライドsで計算をすると式(2)となる．ストライドとは入力をフィルタに通すときにずらす間隔のことである．

$$u_{ij} = \sum_{p=0}^{H-1} \sum_{q=0}^{H-1} x_{si+p,sj+q} h_{pq} \quad (2)$$

出力後の次元数zはフィルタ枚数nとしたとき式(3)となる．

$$z = ((I - (P-1)) \times (J - (Q-1))) \times n \quad (3)$$

Deep Q-Network では複数のフィルタによる畳み込み層を複数回行うため，特徴抽出後の次元数が膨大になり過ぎてしまう．入力の次元数に応じて全結合層の規模を大きくする必要があるため，前向き及び重み更新の計算の処理が重くなると考えられる．また，勾配計算により畳み込みのフィルタの計算も行うため，全体の層数が増えると出力層から遠い勾配が小さくなり，学習完了までの試行回数が増加すると考えられる．

2.2. Dynamic Q-Network

Dynamic Q-Network(SOINN-QN)[5]とは，畳み込みを用いた Deep Q-Network(CNN-QN)の特徴抽出部分に SOINN を採用し，動的に特徴抽出後の次元数を増減しながら動的に学習する手法である．SOINN による特徴抽出では，Q-Network の勾配による重みの更新と同時に独立して競合層のノードの更新及びノイズノードの削除を行うことが可能である．図 2 に Dynamic Q-Network のモデルを示す．

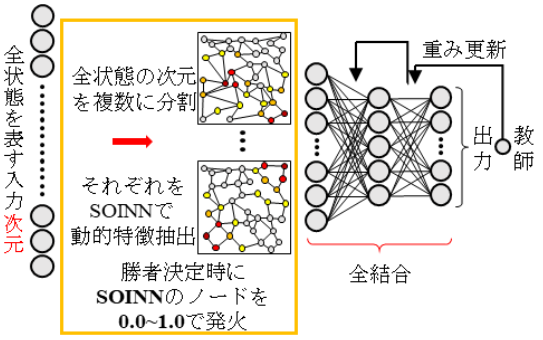

図 2 Dynamic Q-Network のモデル

特徴抽出には一定の数に分割した SOINN を使用する．分割した SOINN はそれぞれ独立してノード数の調整やノードの更新を行う．

Q 値の計算には，まず入力から複数の SOINN の勝者ノードを決定する．そして全結合層のi番目の入力N_iを入力Xと$C1$とのユークリッド距離$dis(X, C1)$用いて式(4)に従い決定する．

$$N_i = dis_{(X,C1)} / dis_{(X,i)} \quad (4)$$

SOINN の入力から 0.0~1.0 に全結合の入力が決定し，類似度が近いほど入力が 1.0 に近くなる．提案手法では，勾配に依存しない環境の学習により従来よりも学習効率の向上が期待できる．また，特徴抽出後の次元数を自動で削減することにより全結合層の処理速度の向上が期待できる．

3. 提案手法

本研究では，特徴抽出部の SOINN と強化学習を行う Q-Network の融合に向けて，２つの改良手法を提案する．

3.1. 削除ノードのネットワークの重みの処理

SOINN の学習によってノードの削除が行われる場合，削除ノードに付随する全結合ネットワークの重みの削除も行う．全結合層の重みはすべてのノードのバランスによって成り立っているため，1 つのノードにつながる重みを完全に削除してしまうと全体のバランスが崩れてしまう．そこで動的に特徴抽出を行う SOINN 部のノード削除時において，Q-Network で結合荷重のコサイン類似度が最も近いノードの重みに影響を残すことで，ノード削除による学習率の低下を防ぐ．図 3 にこの手法のイメージ図を示す．

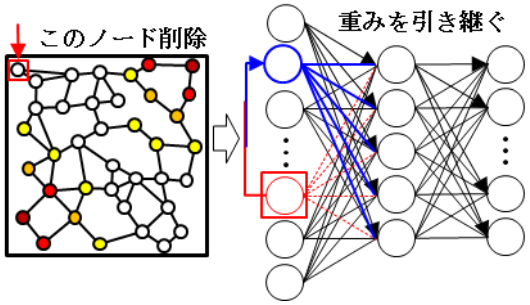

図 3 ノード削除時の重みを引き継ぐ例

SOINN ノードと全結合層の入力ノードはそれぞれ紐づけられている．SOINN 部の赤枠のノードを削除する場合，全結合層の赤ノード及び赤線の重みも消去される．その際に赤ノードと最も重みの類似度が高い青ノードへ赤ノードの重みを引き継いでいる．

3.2. Q-Networkによる状態空間削除の提案

従来の Dynamic Q-Network ではノードの削除判定を特徴抽出部の SOINN のみで行っていた．そこで全結合層の重みの情報を用いて Q-Network 側からも不必要な状態空間ノードの提案を行うことで双方向のやり取りが可能なのではないかと考えた．まず Q-Network で学習により重要度が低下した，あるいは冗長となったと考えられるノードを選出する．各ノードの重みの絶対値総和を求め，最の小さいノードが選出される．そして該当したノードを SOINN の削除候補に取り入れることで実現する．SOINN 単体だけでなく，Q-Network 側からも不必要な状態空間を検出することで，計算量の削減と過学習の防止を目指す．図 4 に提案手法のイメージ図を示す．

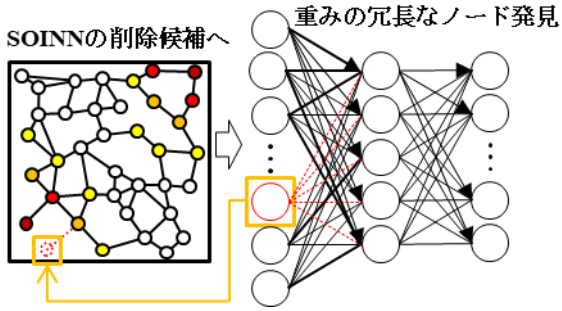

図 4 削除ノード提案のイメージ図

全結合層の重みはそれぞれ影響度が異なる．全結合層内で影響度の少ない赤ノードを発見する．その後，発見した赤ノードに紐づいている SOINN 部のノードを SOINN 部の削除基準に反映させ，ノードの削除を決定している．

4. 実験

実験環境には Reversi を用いる．先攻(黒)と後攻(白)の両方を Q-Network 自身の対戦で学習を行う．自身の対戦で勝利したサイドのみをランダムに打つ相手と戦わせて勝敗を記録する．

本研究では学習にあたって Reversi の盤面を 4 種類8×8の入力データとして変換し盤面の入力データとして学習に用いている．通常の Reversi の盤面を図 5 に示す．

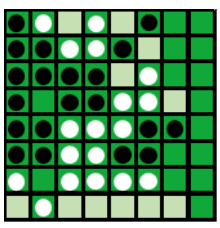

図 5 黒のターンの Reversi の盤面

8×8 の盤面には各マスに「黒の石・白の石・何も置かれていない」の 3 種類の状態がある．Figure の薄緑のマスは現在のターンの黒が石を打つことが可能なマスである．この通常の盤面から 4 種類の盤面情報に抜き出した構成を図 6~9 に示す．確定石とは相手の石によって試合内で裏返ることがない自分の石のことを指す．

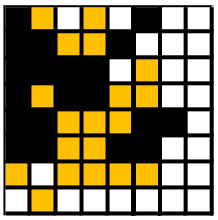

図 6 白黒盤面情報

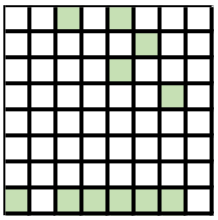

図 7 打てる場所

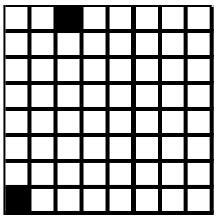

図 8 増加確定石

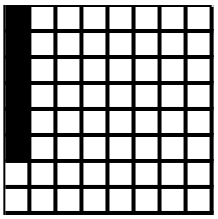

図 9 白黒確定石盤面

図6) 自分の石が置かれていると1.0，相手の石が置かれていると−1.0，何も置かれていないと0.0を入力とする．

図7) 自分の石が置くことが可能だと1.0，置くことができないと0.0を入力とする．

図8) 自分の石を置くことで自分の確定石が増えると1.0，一手打った後の相手の手で相手の確定石が増える手があると−1.0，どちらにも該当しないと0.0を入力とする．

図9) 自分の確定石が置かれていると1.0，相手の確定石が置かれていると−1.0，どちらでもないと0.0を入力とする．

本実験で用いた Dynamic Q-Network の特徴抽出のモデルを図 10 に示す．

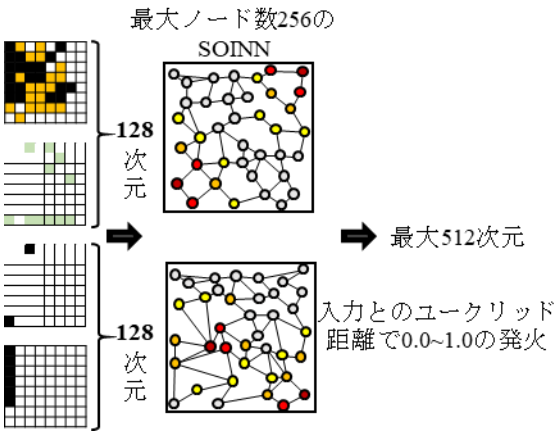

図 10 Dynamic Q-Network のモデル

入力は同様に 4 種類の盤面情報をすべて足した 256 次元とする．4 種類の盤面を 2 種類ずつに分割してそれぞれ独立した SOINN により学習を行う．1 個の SOINN の最大ノード数を 256 個とし，学習により動的に特徴抽出後の次元数を変化させながら学習を行う．全結合層の入力次元数は最大 512 次元となる．

5. 結果

提案手法の削除ノードのネットワークの重みの処理を p1，Q-Network による状態空間削除のアプローチを p2，両方の提案手法を加えたものを p1+p2 として実験結果の比較を行う．図 11 にランダムに打つ相手との勝率を示し，表 1 に処理時間及び p2 の削除確率を示す．1Episode につき，1 回ランダムと戦わせて勝敗を記録し，各 Episode 時点での過去 100 回の勝敗の平均を勝率として計算を行った．

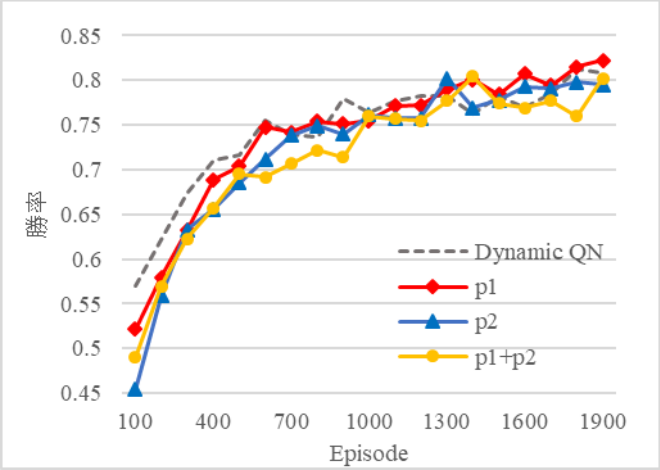

図 11 ランダムに打つ相手との勝率

表 1 学習にかかる処理時間及び p2 の削除比率

	処理時間(s)	p2による提案ノード削除確率
Dynamic QN	**622.4**	
p1	648.8	
p2	623.0	0.014
p1+p2	639.0	0.011

最終的な勝率の高さが最も高くなったのはコサイン類似度により削除ノードの重みの影響を結合加重の近いノードに残す p1 であった．また，従来の Dynamic Q-Network よりも今回の 2 種類の提案手法のいずれかを加えた方が初期の学習の早さ・最終的な勝率が優れていた．2 種類の提案手法を合わせた p1+p2 が最も学習精度が高くなることに期待したが，現状はそこまで至っていない結果となった．p2 では今回重みの絶対値が最も小さいノードを削除条件としていたが，重みの更新時にスパースを取り入れることでより本来の目的である 0.0 に限りなく近い不必要なノードを発見できると考える．

6. まとめ

本研究では，特徴抽出部の SOINN と強化学習を行う Q-Network の融合に向けて，2 つの改良手法を提案した．まず動的に特徴抽出を行う SOINN 部でノード削除時において，Q-Network で結合荷重のコサイン類似度が最も近いノードの重みに影響を残すことで，ノード削除による学習率の低下を防ぐ．また，Q-Network で学習により重要度が低下したと考えられるノードを選出し，SOINN 空間のノードの削除候補に取り入れることで，計算量の削減と過学習の防止を目指した．

実験には Reversi を用いてランダムに打つ相手との勝率及び処理時間で提案手法の有効性の評価を行った．従来の Dynamic Q-Network よりもどちらか一方でも提案手法を加えた方が学習精度の向上が確認できたが，両方とも取り入れることで更なる制度の向上を得ることはできなかった．また，全結合層によって提案された削除ノードが SOINN 部側で採用される確率が非常に低いため改善が必要である．

これらの結果により，Dynamic Q-Network ではノードの削除時に重みのバランスを考慮することで学習率の低下を防ぐことが可能なことが確認できた．また，全結合層の重みの絶対値を条件に削除ノードを選定する場合，スパースモデリングを適用することでより理想的な条件を示せるのではないかと考える．今後の課題として，Dynamic Q-Network を Reversi とは別の環境に適用させ，学習の向上の期待ができるのかを検討していく．

参考文献

[1] Sutton, R. & Barto A., “Reinforcement Learning: An Introduction”, MIT Press, 1998
[2] Volodymyr Mnih et al., “Human-level control through deep reinforcement learning”, Nature 14236, 2015
[3] Y.LeCun et al., “Backpropagation Applied to Handwritten Zip Code Recognition”, Neural Computation 1, pp.541-551, 1989
[4] 山崎和博 他, “自己増殖型ニューラルネットワーク SOINN とその実践”, 日本神経回路学会誌, Vol.17, No.4, pp.187-196, 2010
[5] 小松泰士，山内ゆかり, “競合学習で特徴抽出を行う動的強化学習ネットワーク”, 2020