リカレントネットを用いた強化学習による記憶を利用した探索行動の学習

後藤健太、 柴田克成
(大分大学 工学部)

1 はじめに

強化学習[1]では、行動選択に導入した乱数要素を一般的に探索と呼び、この探索による試行錯誤を繰り返すことによって、限られた情報から自ら学習することができる。一方、われわれが通常、探索行動と呼ぶものは、乱数を用いた確率的なものではなく、過去の情報などの知識を用いた知的で決定論的なものであると考えられる。この点で注目してわれわれは、そのような探索行動を、強化学習によってエージェントが自律的に獲得できるかどうかを調べてきた。

先行研究[2]では、記憶を用いた自律的な学習を行うために、情報の保持を可能とするリカレント構造をもったニューラルネットを強化学習に組み込んで学習を行った。その結果、簡単な迷路問題の学習において、探索行動がある程度獲得できることが分かった。しかし、条件によってはうまく学習できない場合があることが報告されていた。

そこで、本稿では先行研究と同一のタスクにおいて、学習後の行動や行動価値をより詳細に検討した結果を報告する。

2 学習方法

本研究では、強化学習のアルゴリズムとしてQ-learning[3]を用い、リカレントニューラルネットとして中間層ニューロンの出力を次の時刻の入力に付加する Elman 型のニューラルネットを用いた。そして、Q-learning のアルゴリズムに基づいて、ニューラルネットの教師信号を求め、教師あり学習を行うことで、強化学習によってニューラルネットを学習する。

まず、現在の状態をニューラルネットに入力する。ニューラルネットの出力ニューロンは、各行動の数と同じだけ用意し、その出力を各行動の行動価値、つまりQ値として扱う。そして、各行動のQ値に小さめの乱数を足して最大の値となる行動を選択する。その際に、足される乱数の幅は学習回数と共に小さくしていき、学習終了時にはほぼ0とした。

学習時は、選択した行動a_tに該当する出力のみに教師信号を与えて学習させる。現在の状態 s_t における行動 a_t の教師信号 $T_{at,t}$ は、その行動をした後の状態における観測値 s_{t+1} をニューラルネットに入力したときに最大となるQ値を使って次式のように自動作成する。

$$T_{a_t,t} = r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') \quad (1)$$

r_t は報酬を表し、γ は割引率を表す。γ の値の範囲は($0 \leqq \gamma \leqq 1$)で与えられる。この教師信号を用い、BPTT 法[4]に基づいて時間をさかのぼって教師あり学習を行う。このQ値は

$$Q^*(s_t, a) = E^*\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s, a_t = a\} \quad (2)$$

となるように定義され、(1)式はこの値を逐次的に求めるための式となっている。E^*はそれ以降の最適な行動をとった場合の期待値を表し、k は、報酬を得るまでのステップ数を表している。この式より、基本的には、報酬が得られるまでのステップ数が大きくなるほど Q 値は小さくなり、報酬の可能性が確率的に表される場合にも、期待値計算により、Q 値の理想値を算出することができる。また、リカレントネットを用いることで、過去の観測値を保持しなくても、リカレントネットが学習を通して、必要な情報を抽出して中間層に保持するようになることを期待する。

3 シミュレーション

本研究では、記憶を利用しなければ効率のよい探索ができないと考えられる簡単な迷路でシミュレーションを行った。迷路は大きさを3×3マスとし、内部に閉じた空間を作らない範囲で毎回ランダムに4つの壁が設置される。ゴール位置も毎回ランダムに決められる。また、ゴールはエージェントには見えないものとする。エージェントの初期位置は常に中心に固定する。

エージェントに与えられる情報は、周囲の上下左右の壁の有無と、直前の行動が 4 つの各方向への行動のうちどれだったかを示すものの計 8 つである。したがって、エージェントは、現在の観測値からだけでは、迷路の全体の形状を認識したり、ゴールの位置を推定することはできない。よって、このタスクでは、過去に訪れた状態を適確に記憶し、無駄な行動をすることなく、見えないゴールに早く到達するように効率的に探索することが求められる。

1 試行中で迷路内の探索が進み、通っていないマスの数が減ると、次以降に未探索のマスに行く際に、そこがゴールである確率が大きくなる。したがって、一般的に Q 値は探索が進むにつれて大きくなる。たとえば、Fig.1の迷路で点線のような経路で行動したとき、ゴール直前の状態では、次に必ずゴールすることができるため、理想的なQ値は報酬 r の値と一致する。しかし、途中の左下に行った状態では行き止まりであり、既に探索したマスを戻る間は報酬が得られないので、Q値は一旦下がる。しかしながら、3×3の迷路であることや、一度通ったところどうかの区別が学習によって理解できるようになっていなければ、このような考え方による適切な Q 値は得られない。

Fig.2に学習後のエージェントがFig.1の迷路で点線のような経路をとった場合の各状態での最大 Q 値の推移と式(2)の期待値計算に従った理想値の変化を示す。また比較として、リカレント構造をもたない通常のニューラルネットを導入した場合のQ値の変化を示す。

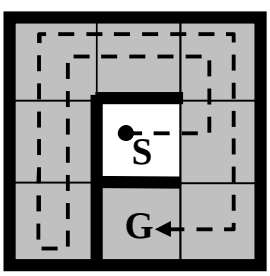

Fig.1 An example of agent's behavior after learning.

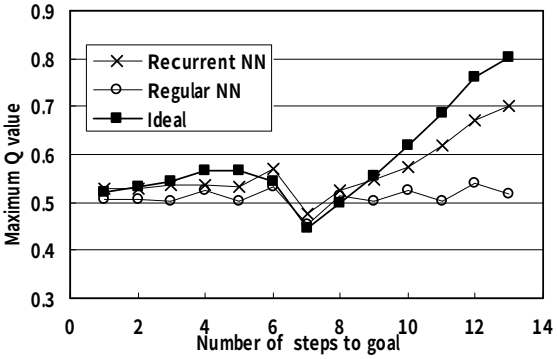

Fig.2 Comparison of maximum Q values among ideal and actual outputs derived by neural networks.

通常のニューラルネットを導入した場合でも最適な探索行動を学習することができた。しかし、Fig.2 の最大 Q 値の変化を見ると、(0,0)の行き止まりのところで少し下がっている以外は、最後まで同じような Q 値が続いている。過去の情報を記憶できないため、常に同じ確率で次に移動するマスにゴールが存在することを予測した評価を行っていると考えられる。

一方、リカレントネットを導入したものについては、しばらくはゴールが存在しない折り返し地点直後で評価値は落ちて、ゴールが限定される終盤の Q 値は上昇しており、全体として、理想値と近い傾向を捉えているように思われる。

この問題は一見簡単に見えるが、それはあくまで、われわれが迷路というものをイメージできるからある。しかし、エージェントには、そのようなイメージを作る知識はなく、壁の状態と直前の行動が単なる時系列として与えられているだけである。人間にとっても、単に 8 個数字の羅列である入力が毎ステップ与えられる問題だとすれば、問題自体が難しいものであるとも考えられるとされた。そして、それが理想値と完全に一致しない理由ではないかと考えている。

次に、先行研究で最適な探索行動を学習できなかったタスクである、ゴールの存在位置を四隅のみに限定した場合でのシミュレーション結果を示す。

このタスクはリカレント構造を持たない通常のニューラルネットでは学習できなかった。しかし、リカレントネットを使って再度学習を試みたところ、学習係数などのパラメータを調整し、学習に時間を費やせば、ごく一部の形状の迷路を除いては、最適な探索行動を学習することができ、また探索の進行にともなう Q 値の上昇も確認できた。最適な行動をとれなかった場合の一例を Fig.3 に示す。

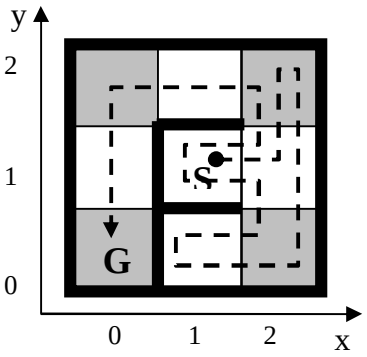

Fig.3. Agent's redundant behavior after learning when a goal position is limited to one of the four corners.

エージェントは(2,2)へ移動した後、3 ステップ目に、下向きの行動を選択している。これは一見最適な行動ではないように見えるが、式(2)の期待値計算を用いると、最適な行動であることが分かる。さらに進んで、8 ステップ目には、(2,1)から左へ向かう行動を選択している。(1,1)は初期位置であり、最初の状態で周囲の壁の位置を確認できているため、この場所を通過して、他の未探索の場所に行けることはなく、かつゴールがこの場所に出現することはあり得ない。

このような最適でない探索行動をとってしまった原因としては、エージェントは、(2,1)を通った後は(2,2)にゴールがないことを確認した上で下に降りているため、さらにその前に(1,1)から(2,1)に来たということを忘れてしまったためと考えられる。つまり、記憶すべき新しい情報が入ってくることで、古い情報を忘れてしまい、既に行ったはずの(1,1)に向かうような行動をとってしまったことが考えられる。

他の最適な行動がとれなかった場合を、いくつかピックアップして調べたところ、同じような傾向が観察された。

4 まとめ

強化学習によって、ある程度の乱数によらない行動を獲得することができ、またリカレントの導入により、ある程度正しく現在の状態を評価し、それを利用した効率的な探索を行えていることを示すことができた。しかし、多段階の記憶が必要なタスクの学習は困難であり、それについてはリカレントネットの記憶の学習能力の問題が考えられる。より効率的な探索を学習するためには、リカレントネットの学習能力の向上が必要と考えられる。

謝辞

本研究は、日本学術振興会科学研究費補助会 #19300070 によって補助された。

参考文献

1) Sutton, R.S. and Barto, A.G., "Reinforcement Learning:An Introduction",MIT Press, Cambridge, MA(1998)

2) 柴田克成、"強化学習による探索行動の学習"、システム・情報部門学術講演会議演論文集、pp11-16、計測自動制御学会(2005.11.28-30)

3) C.J.C.H Watkins and P. Dayan, "Q-earning", Machine Learning, Vol.8, pp.279-292, 1992.

4) Rumelhart, D.E., Hinton, G.E., and, Williams, R.J., "Learning Internal Representations by Error Propagation". Parallel Distributed Processing, The MIT Press, pp. 318-362(1986)