人工知能学会研究会資料
SIG-AGI-008-08

深層学習とモンテカルロ木探索を用いた強化学習の組合せ最適化問題での実験

Experiments on Combinatorial Optimization with Reinforcement Learning Using Deep Learning and Monte Carlo Tree Search

疋田 聡

Satoshi Hikida

Abstract: Reinforcement learning using deep learning and Monte Carlo tree search has been reported to be extremely effective as an artificial intelligence algorithm that is used in AlphaZero etc. and is widely applicable to various games. Since this method is essentially an algorithm that solves the search problem efficiently, it is possible to solve a general combination optimization problem as well as a game. Therefore, in order to deepen the understanding of this method, experiments were applied to combinatorial optimization problem, and the results are reported.

1. 背景

2016年の3月にAlphaGo[3][6]がその当時人類最強と言われる囲碁棋士の一人を破り社会に衝撃を与えた。その後、AlphaGo では初期学習のデータに人間の打った棋譜データを用いていたが、2017 年 10 月に発表された AlphaGo Zero[5]では、人間の打った棋譜データをまったく用いずに、0 から学習して以前の AlphaGo よりも強くなったということで驚かされた。さらに、2017 年 12 月に発表された AlphaZero[4]では、AlphaGo Zero と同じ深層学習とモンテカルロ木探索を用いた強化学習を用いて、ゲームのルールを変更するだけで、チェスや将棋で最強レベルの強さを達成することが可能であることが報告された。

このように、深層学習とモンテカルロ木探索を用いた強化学習は、様々なゲームに汎用的に適用可能な人工知能アルゴリズムとして非常に有効である。またこの方法は、本質的には探索問題を効率的に解くアルゴリズムなので、ゲームだけでなく一般的な組合せ最適化問題を解くことも可能であると考えられる。

そこで、この方法の理解を深めるため、まず深層学習とモンテカルロ木探索を用いた強化学習について説明し、その後に一般的な組合せ最適化問題に適用する実験について説明する。

2. 深層学習とモンテカルロ木探索を用いた強化学習

深層学習とモンテカルロ木探索を用いた強化学習について、AlphaGo Zero を例として説明する。AlphaGo Zero では図 1 のように、囲碁の盤面の石の配置の状態 s をニューラルネットの入力とし、モンテカルロ木探索で方策πによって行動を選択し、行動結果から得られる報酬 V を用いて状態 s から予測報酬 v と方策πによる行動確率 p を出力するニューラルネットの学習を、式 1 を用いて行う。

$$(\boldsymbol{p}, v) = f_\theta(s) \text{ and } l = (z-v)^2 - \boldsymbol{\pi}^{\mathrm{T}} \log \boldsymbol{p} + c\|\theta\|^2$$

式 1　損失関数(文献[5]の式(1)より引用)

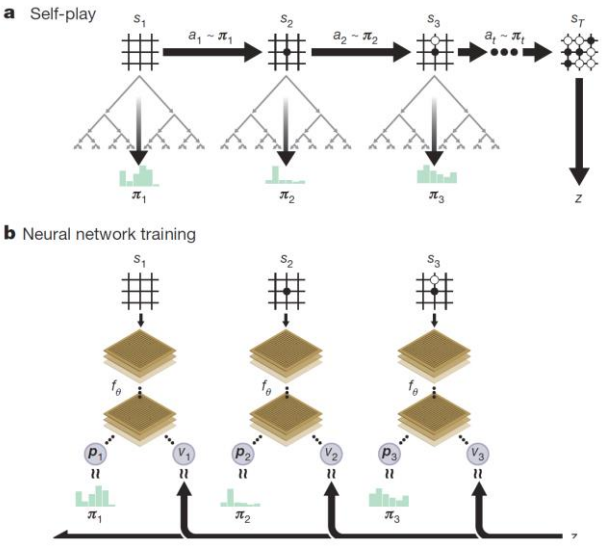

図 1　AlphaGo Zero での深層学習とモンテカルロ木探索を用いた強化学習(文献[5]の Figure 1 より引用)

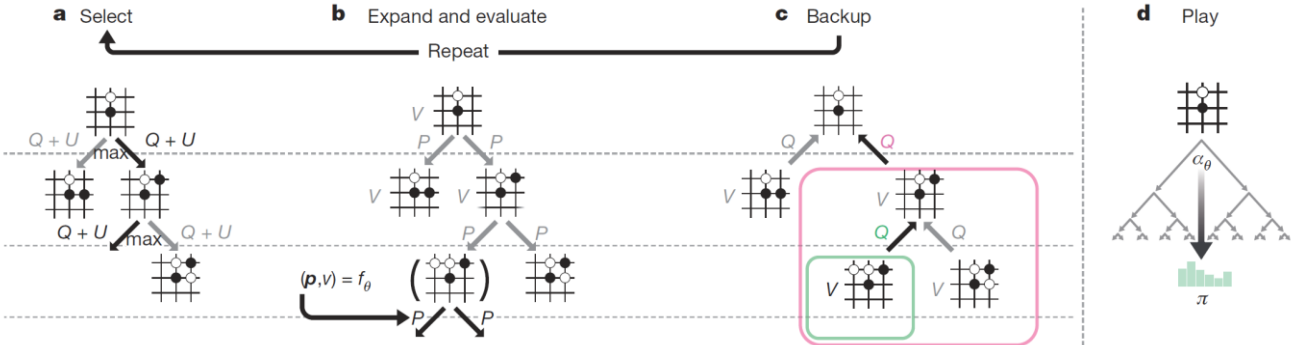

図 2　AlphaGo Zero でのモンテカルロ木探索(文献[5]の Figure 2 より引用)

報酬 V はモンテカルロ木の末端のリーフノードでゲームの勝敗から得られ、図 2 のように報酬を上位ノードへ伝搬させ、訪問回数で割った Q と式 3 で算出した U を用いて式 2 で行動を選択してモンテカルロ木の探索を行っている。

$$a_t = \underset{a}{\operatorname{argmax}}(Q(s_t, a) + U(s_t, a))$$

式 2　モンテカルロ木探索での行動の選択(文献[5]の Methods より引用)

$$U(s,a) = c_{\text{puct}} P(s,a) \frac{\sqrt{\sum_b N(s,b)}}{1 + N(s,a)}$$

式 3　モンテカルロ木探索での U 値の算出(文献[5]の Methods より引用)

このように、深層学習とモンテカルロ木探索を用いた強化学習では、深層学習を用いた盤面パターンからの直観的な報酬予測と、モンテカルロ木探索による決定的な先読みを組み合わせることにより、効率的な探索を実現している。

3. 組合せ最適化問題への適用

深層学習とモンテカルロ木探索を用いた強化学習は、本質的には探索問題を効率的に解くアルゴリズムなので、ゲームだけでなく一般的な組合せ最適化問題を解くことも可能であると考えられる。

そこで、この方法の理解を深めるため、一般的な組合せ最適化問題の一つである巡回セールスマン問題[1]に適用する実験について説明する。

巡回セールスマン問題は古くからある有名な組合せ最適化問題の一つであり、与えられた都市を全て 1 回ずつ訪れたときの経路距離の総和が最小になる経路を探索する問題で、計算複雑性理論において NP 困難と呼ばれる問題のクラスに属する。また、巡回セールスマン問題のベンチマーク問題集として、TSPLIB[2]などが公開されている。

巡回セールスマン問題の問題例を図 3 に示す。

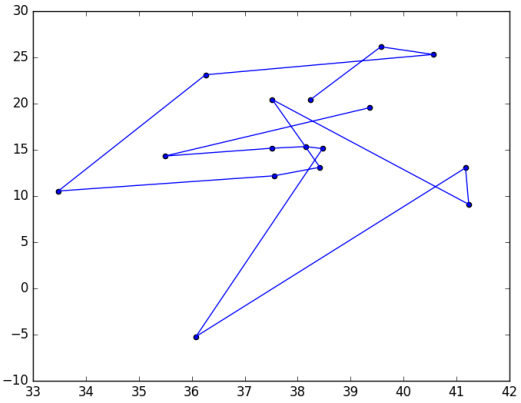

図 3　巡回セールスマン問題の例

深層学習とモンテカルロ木探索を用いた強化学習では、前章で説明したように、深層学習を用いた図形パターンからの直観的な報酬予測と、モンテカルロ木探索による決定的な先読みを組み合わせることにより、効率的な探索を実現していると考えられるが、巡回セールスマン問題でも都市の配置に図形的なパターンがあるので、実験がアルゴリズムの性質の理解に役立つのではないかと考えられる。

3.1. ニューラルネットへの入力形式への変形

ニューラルネットへ巡回セールスマン問題の状態を入力するため、都市の配置をある程度反映してニューラルネット上に配置し、訪れた都市を削除してゆくという形で状態を表現する。図 4 に例を示したように、残り都市数が 5 個の上の状態から 5 行 2 列目の都市を訪れた場合、下の図のようにその都市を

削除し、残り都市数が 4 個になる。

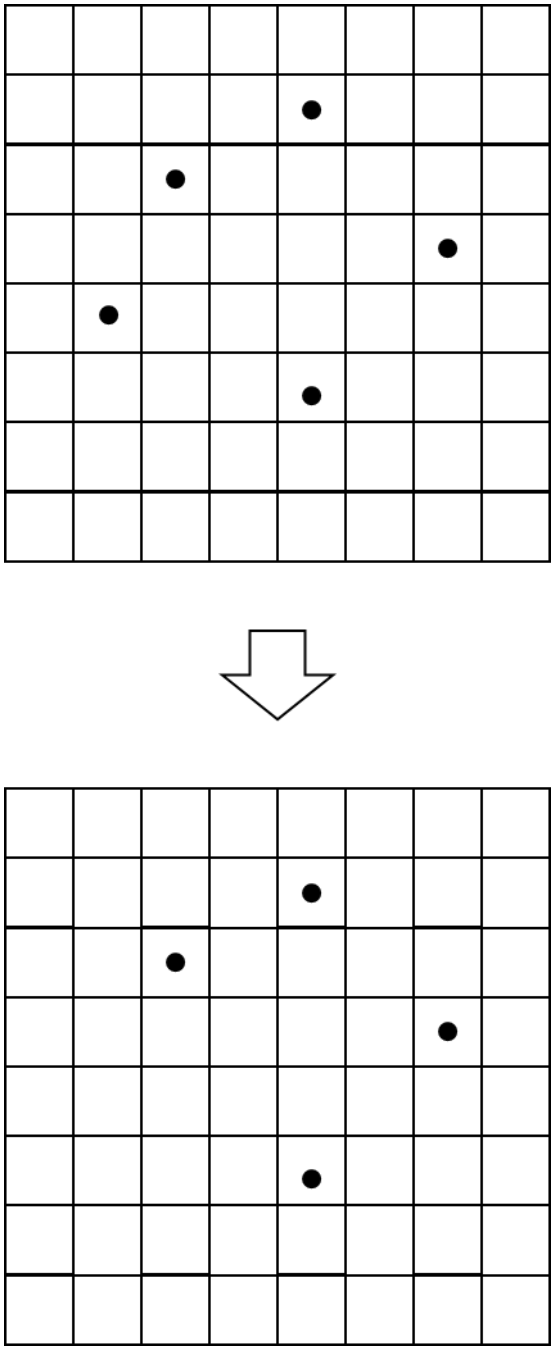

図 4　ニューラルネットへの入力形式の例

また、上記で都市の配置を「ある程度」反映してという書き方をしたのは、ニューラルネットは都市の配置の図形的なパターンから直観的な報酬予測を行う役割であり、より正確な報酬値はモンテカルロ木探索による決定的な先読みで、都市の本当の座標値データから求めているので、都市の配置を「ある程度」反映していれば少々誤差があっても直観的な報酬予測が可能であると考えられるからである。これにより、都市の座標からニューラルネットへの入力形式に変換したときに、都市が同じ点に重なってしまった場合に、隣の空いている点にずらすなどの柔軟な対応が可能となる。

3.2. 現在の位置情報と始点情報の追加

上記のように使用していない残りの都市の配置だけをニューラルネットへの入力とすると、次に回る都市を選択したときに現在どの都市にいるのかが分からないため、距離の増加量が算出できないという問題がある。また、最後の都市まで回ったときに、最後の都市と最初の都市を結んだ距離も経路距離の総和に加える必要があるが、始点がどこだか分からないと距離の算出ができないという問題もある。

そこで、現在状態に加えて、前回の都市位置、最初の都市位置を別のチャンネルに追加し、3 チャンネルのデータをニューラルネットへの入力としたことで、これらの問題へ対応した。

3.3. 報酬の定義と報酬伝搬方法の修正

巡回セールスマン問題の報酬は経路距離の総和と考えられるが、距離が短い方を探索したいので、報酬は経路距離の総和で符号を逆にしたものとする。また、囲碁などのゲームでは勝敗を報酬とすると最終局面が同じであれば報酬は同じになるが、巡回セールスマン問題で今回のようなニューラルネットへの入力形式を用いる場合は、最終状態が同じでもそこまでの経路が異なると報酬が異なってしまうという問題がある。

そこで、囲碁などのゲームとは異なり、終了状態から逆向きに経路長を足しながら進んで、終了状態からそのノードまでの経路距離の総和から報酬を計算することで、最終状態が同じでもそこまでの経路が異なると報酬が異なってしまうという問題を解決した。具体的には、図 2 でリーフノードからルートノードに V を伝搬させる各段階で、そのノードでの経路距離を反映させるようにする。

また、囲碁などのゲームでは、報酬が勝ち負けで 1,-1 なのに対し、巡回セールスマン問題の経路距離は 1,-1 と比較して大きな値や小さな値になることに注意が必要である。そこで、試行で得た学習データから経路距離の総和の最良値と中央値を求め、

V’ = (V - 中央値) / (最良値 - 中央値)

という正規化を行う。

3.4. 対戦相手の省略

ゲームでは対戦相手が存在するためプレイヤーが交互に手を打つ形になるが、組合せ最適化問題では対戦相手は存在しないため、対戦相手の手は全てパスしたものとして処理する。

4. 実験方法

実験環境として、ハードウェアは CPU：Intel Core i5 8400、GPU：Nvidia GTX 1080ti、OS は Ubuntu16.04LTS、ソフトウェアは Python、Tensorflow、Keras を用いている。

5. 実験結果

深層学習とモンテカルロ木探索を用いた強化学習の組合せの最適化問題への適用の例として、上記で説明した方法により巡回セールスマン問題へ適用した実験を行っているが、学習時間の関係上この予稿には間に合わないため、実験結果について研究会の発表時に提示する予定である。

6. まとめと今後

AlphaZero 等で用いられ様々なゲームに汎用的に適用可能な人工知能アルゴリズムである深層学習とモンテカルロ木探索を用いた強化学習を一般的な組合せ最適化問題の一つである巡回セールスマン問題に適用する実験方法についての説明し、アルゴリズム修正により適用可能であることを示した。

今後は、上記で説明した方法を用いた実験を続行し、実験結果の解析と考察を行っていく予定である。

参考文献

[1] Dijkstra, E. W.: A note on two problems in connexion with graphs, Numerische mathematik, 1(1), 269-271, (1959)

[2] Reinelt, G.: TSPLIB—A traveling salesman problem library, ORSA journal on computing 3(4) 376-384, (1991)

[3] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Dieleman, S.: Mastering the game of Go with deep neural networks and tree search, nature, 529(7587), 484-489, (2016)

[4] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., ... & Lillicrap, T.: Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm, arXiv preprint arXiv:1712.01815, (2017)

[5] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Chen, Y.: Mastering the game of go without human knowledge, Nature, 550(7676), 354, (2017)

[6] 大槻知史, 三宅陽一郎 監修: 最強囲碁 AI アルファ碁 解体新書 深層学習、モンテカルロ木探索、強化学習から見たその仕組み, 翔泳社, (2017)