人工知能学会研究会資料
SIG-AGI-009-07

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とフレーム問題に関する１考察

Experiments on Combinatorial Optimization with Reinforcement Learning Using Deep Learning and Monte Carlo Tree Search and a Consideration of Frame Problem

疋田 聡 [1]

Satoshi Hikida [1]

[1] AGICRON 研究所株式会社
[1] AGICRON Research Institute, Ltd.

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. The relationship between this method and the frame problem also be described.

1. 背景

2016 年の 3 月に AlphaGo[10][13][14]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[12][14]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[11]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。

さらに、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であるが、この方法はゲームのみならずフレーム問題への対応にも関係していると考えられるので、それについての考察も加えて述べる。

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、モンテカルロ木探索を行っており、終了局面以外のモンテカルロ木の探索の先端ノードでは、囲碁の盤面の石の配置の状態 s を入力としたニューラルネットによって、方策 π と報酬 V の近似値を求め、その値を利用して行動を選択し、自己対戦によって学習データを作成する。また、自己対戦データを学習データとして用いて、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[12]の式(1)より引用)

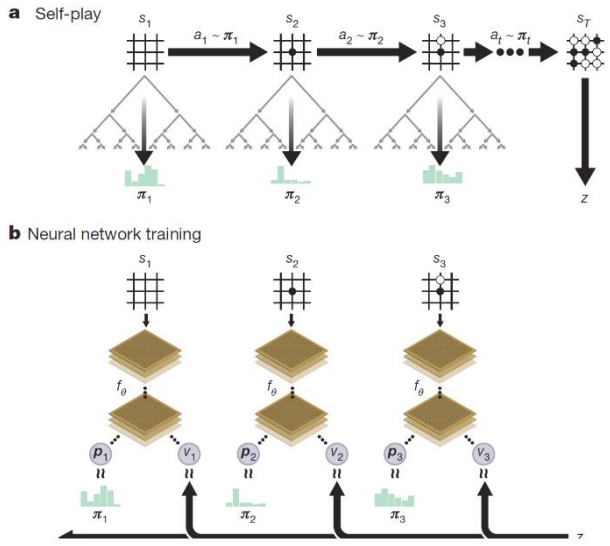

図 1　AlphaGo Zero での深層学習とモンテカルロ木探索を用いた強化学習(文献[12]の Figure 1 より引用)

報酬 V はモンテカルロ木の末端のリーフノード（終了局面）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[12]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[12]の Methods より引用)

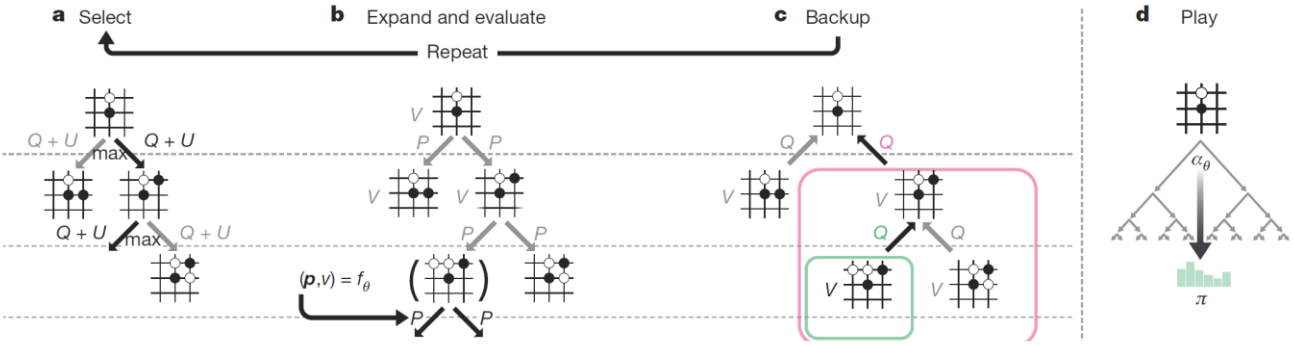

図 2　AlphaGo Zero でのモンテカルロ木探索(文献[12]の Figure 2 より引用)

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[3]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[9]などが公開されている。

巡回セールスマン問題の問題例として、TSPLIB の”eil51”で先頭から 21 都市を抽出し、分枝限定法[6]を用いて全探索した結果を図 3 に示す。

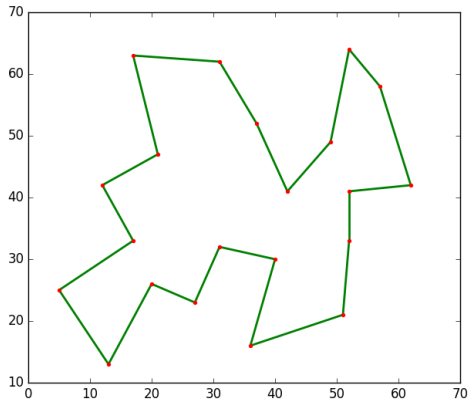

図 3　巡回セールスマン問題の例

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。

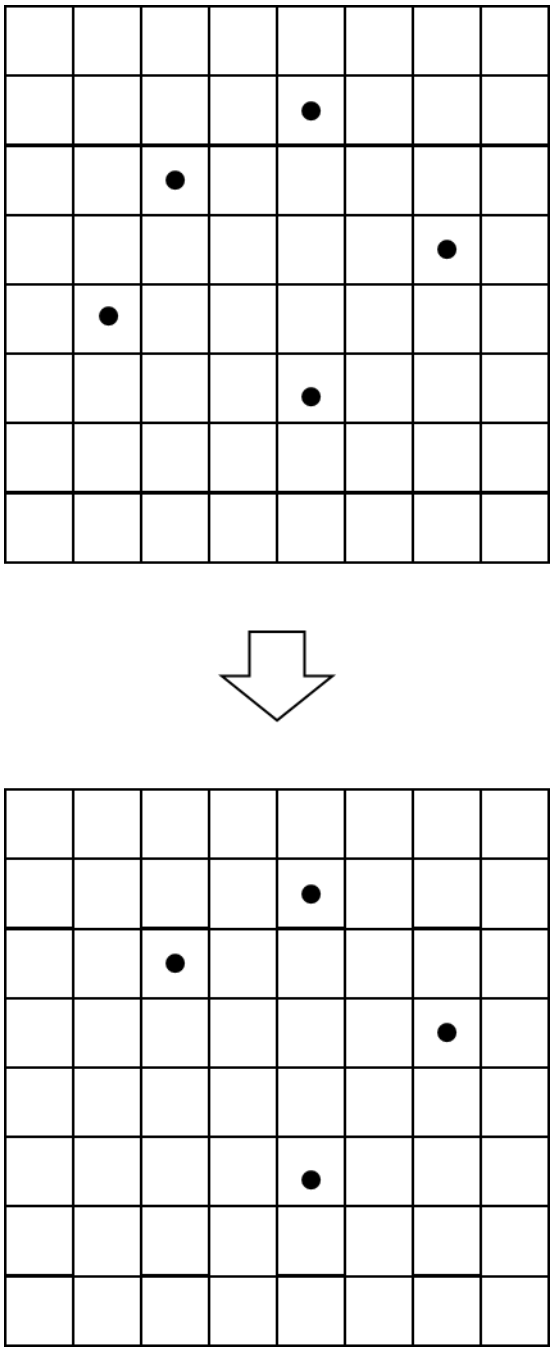

図 4　ニューラルネットへの入力形式の例

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、正規化を行う必要があるが、前回の報告[16]では、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)

という正規化を行ったが、正規化の基準値が変動すると解析が行い難いので、今回は、基準値をクリストフィードのアルゴリズム[1][6]で得られる値、下限値を最小 1-木とクリストフィードのアルゴリズムで得られた値を 1.5 で割ったものとの大きい方を用いて、下記の式で正規化した。

V’ = (V – 基準値) / (下限値 – 基準値)

なお、クリストフィードのアルゴリズムは巡回セールスマン問題において多項式時間で最適解の 1.5 倍以下になることが保証されている近似アルゴリズムである。

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。

3.5. 好奇心レベルの導入

ゲームでは対戦相手が着手を変えてくるため、局所的に良い評価値が得られたときに探索が進まなくなるという状態にはなり難いが、相手のいない組合せ最適化問題の探索では、局所的に良い評価値が得られたときに探索が進まなくなることが考えられる。このような状態に陥るのを防ぐため好奇心[15]を利用することが考えられる。そこで、終了局面以外の探索の先端ノードで用いるニューラルネットからの評価値の推定値に好奇心レベルの数値を足すという仕組みを導入した。これにより、まだ終了局面まで達していいないノードが探索されやすくなるという効果が期待される。

3.6. 頻度基準を評価値基準へ変更

囲碁などのゲームや実世界での行動探索では、一回だけ良い評価値が得られても対戦相手のミスや環境ノイズなどがありうるので、頻度基準の方策で行動選択することにより変動要素の影響を軽減することが有効である。しかし、組合せ最適化問題の探索では、対戦相手や環境ノイズなどの変動要素は存在せず、一回だけ得られた良い評価値はそのまま有効であるので、頻度基準の方策から最良値基準の方策の方が探索効率が良くなると考えられる。そこで、今回の実験では AlphaGo Zero とは異なり、頻度基準の方策から最良値基準の方策に変更して実験を行った。また、一度得られたノードの最良値は、次回以降も有効なので、ノードの最良値を保存して再利用できるようにした。

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。

実験対象の巡回セールスマン問題は、TSPLIB[9]の”eil51”で先頭から 21 都市を抽出し、分枝限定法[6]を用いて全探索した結果を図 3 に示す。なお、先頭から 21 都市を抽出しているのは、実験時の詳細な分析のために、分枝限定法を用いて全探索できるように問題の難しさを調整するためである。

5. 実験結果

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結果を図 5,6 に示す。

図 5,6 はアクションステップ数とそれまでに得られたパス長の最短値の関係を示したものであり、6 プロセスの経過をそれぞれ示している。なお、分枝限定法により全探索して求めた最短パス長は 256.12 である。

図 5 は学習前の結果で、最短パス長に達していないことが分かる。それに対し、図 6 は、学習後の結果で、6 プロセスが全て最短パス長に達しているという結果が得られた。

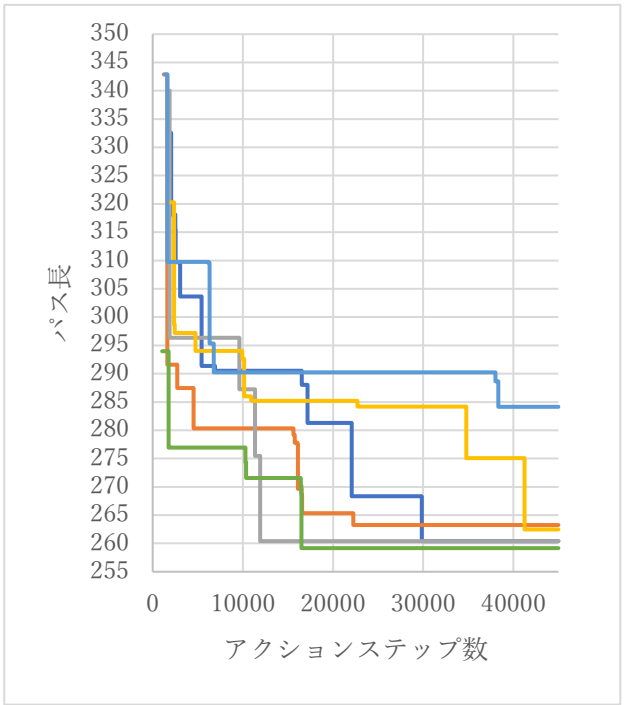

図 5　学習前の実験結果

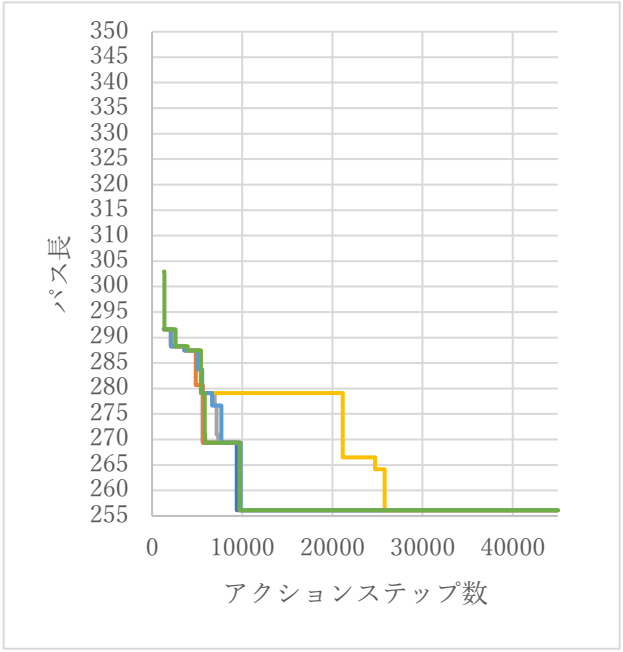

図 6　学習後の実験結果

なお、巡回セールスマン問題に対しては、分枝限定法やタブー探索など、問題固有のヒューリスティックを用いることでより高速に解を得ることも可能であるが、今回は AlphaZero 等のアルゴリズムの理解を深めることを主な目的としたため、問題固有のヒューリスティックは用いなかったが、そのような問題固有のヒューリスティックと深層学習とモンテカルロ木探索を用いた強化学習とを組み合わせてさらに高速化することも可能である。

6. フレーム問題との関係についての考察

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であるが、この方法はゲームのみならずフレーム問題への対応にも関係していると考えられるので、それについて述べる。

6.1. フレーム問題とは

フレーム問題[5][2]は、1969 年に指摘された人工知能の分野の難問の一つである。例えば、ロボットが時限爆弾の仕掛けられた洞窟からバッテリーを取り出してくる課題を考えると、

① バッテリーの上に爆弾が載っている場合には、バッテリーをそのまま持ってくると爆弾

も持ってきてしまうなどの副次的に発生する現象を考慮する必要がある

② しかし、「バッテリーを動かすと上にのった爆弾は爆発しないか」、「爆弾を動かすと天井が落ちてきたりしないか」、「爆弾に近づくと壁の色が変わったりしないか」など副次的に発生しうる無限の可能性を全て考慮するには無限の計算時間が必要になってしまう

③ そこで、「爆弾に近づくと壁の色が変わったりしないか」など目的遂行に無関係な事項は考慮しないようにしようとしても、目的と無関係な事項が無限にあるため、その全て洗い出すための計算時間が無限に必要になってしまう

という問題である。

フレーム問題の説明ではよく、将棋や囲碁などのゲームのような範囲が限定されたタスクでは発生せず、実世界のような範囲が限定されない課題で発生するとなっているが、単純にそう言ってしまっていいのか少し疑問に思ったので、ゲームと対応付けて考えてみる。

6.2. ゲームとの対応付け

まず、フレーム問題の①の副次的に発生する現象を考慮する必要があるというのは、ゲームでは自分の着手に対する相手の着手を考慮する必要があると対応付けることができる。自分の着手に応じて相手が着手を変えてくるのを副次作用と考えることができるからである。

次に、フレーム問題の②の副次的に発生しうる無限の可能性を全て考慮するには無限の計算時間が必要になってしまうというのは、ゲームでは自分の着手に対する相手の着手を全て考慮した先読みを行うと無限に近い計算時間が必要になると対応付けることができる。単純に全探索しようとすると無限に近い時間がかかるからである。

最後に、フレーム問題の③の目的と無関係な事項を全て洗い出すための計算時間が無限に必要になってしまうというのは、ゲームでは勝敗に無関係な相手の着手を全て洗い出すために計算時間が無限に必要になってしまうということと対応付けることができる。評価関数などの事前知識を入れない場合は、勝敗に無関係かどうかも全探索して調べる必要があるからである。

このように対応付けると、ゲームの世界でもフレーム問題と同じ問題が発生しそうであるが、周知のようにゲームでは上記のような問題は起こっていない。以下にゲームでの対応とその対応がフレーム問題へ適用できないかを考えてみる。

6.3. ゲームでの対応とフレーム問題

上記のように、ゲームでも自分の着手に対する相手の着手を全て考慮した先読みを行うと無限に近い計算時間が必要になるが、これは単純に全探索を行うから発生する問題であり、AlphaGo Zero のように、先読みの途中で着手選択の方策や評価値をニューラルネットによって近似し、有望な着手に絞って先読みを行うことにより、実時間で実行可能となっている。また、このニューラルネットは事前知識が 0 でも自己対戦のデータから学習することができ、学習データが全ての局面を網羅していなくてもニューラルネットの汎化能力により未経験局面の方策や評価値も近似できることが重要である。このようなニューラルネットの汎化能力による近似によって人間の能力を超えられるかどうか従来は明らかではなかったが、AlphaGo Zero によって少なくともゲーム分野において人間の能力を超えることが実証されたことには大きな意味があると考えられる。

ここから上記の対応付けを逆に辿って、ゲームでの対応をフレーム問題に適用することを考えると、副次的に発生する現象を全探索のように全て推論しようとするのではなく、実世界での経験または実世界に近いシミュレーションの世界での経験から学習したニューラルネット等の汎化能力によって副次的に発生する現象を近似し、有望な行動に絞って先読みを行うことにより、実時間で実行可能とするという方法が考えられる。すなわち、従来は人間の常識を CYC[4][8]のように人間が論理式で記述しようとしていたが、ゲームでの対応と同様に人工知能が常識のようなものを経験から学習する形になる。ここで、ニューラルネット等の汎化能力によって副次的に発生する現象を近似が十分な精度で行えるのかがポイントとなるが、実世界対応としてロボットへの強化学習の適用[7][17]の進展や、最近 2018 年 8 月に OpenAI の人工知能「OpenAI Five」が複雑な状況認識や判断が必要なゲーム「Dota 2」の 5 対 5 のチーム戦でプロチームに勝ったことなどを見ても、膨大な計算資源や学習データが利用できるようになれば、従来不可能と考えられていたこともできるようになる可能性があると考えられる。

7. まとめ

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に適用する実験について説明し、実験結果を示した。

また、この方法はゲームのみならずフレーム問題への対応にも関係していると考えられるので、ゲームでの対応とフレーム問題についての一つの考察を示した。

参考文献

[1] Christofides, N., Worst-case analysis of a new heuristic for the travelling salesman problem, Report 388 Graduate School of Industrial Administration CMU, (1976)

[2] Dennett, D. C., Cognitive wheels: The frame problem of AI, *Language and Thought*, *3*, 217, (2005)

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

[4] Lenat, D. B., Prakash, M., & Shepherd, M., CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks, *AI magazine 6*(4) 65, (1985)

[5] McCarthy, J., & Hayes, P. J., Some philosophical problems from the standpoint of artificial intelligence, In *Readings in artificial intelligence* (pp. 431-450), (1981)

[6] P.グリッツマン, R.ブランデンベルク, 石田基広(翻訳), 最短経路の本, 丸善出版, (2012)

[7] Peng, X. B., Abbeel, P., Levine, S., & van de Panne, M., DeepMimic: Example-Guided Deep Reinforcement Learning of Physics-Based Character Skills, arXiv preprint arXiv:1804.02717., (2018)

[8] R.デービス, D.B.レナート, 人工知能における知識ベースシステム, 啓学出版, (1991)

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

[10] 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)

[11] 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)

[12] 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)

[13] 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)

[14] 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 増補改訂版 アルファ碁ゼロ対応 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, (2018)

[15] 疋田聡, 好奇心で動機付けされた強化学習の実験, 人工知能学会研究会資料 SIG-AGI-006-09, (2017)

[16] 疋田聡, 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, 人工知能学会研究会資料 SIG-AGI-008-08, (2018)

[17] 疋田聡, 方策最適化による強化学習を用いた人型ロボットの動作学習の実験, 人工知能学会研究会資料 SIG-AGI-007-02, (2017)