機械学習による3次元シミュレーション空間の場面評価手法の検討

石井 徹　矢農 正紀　千葉 優介　岡田 正央　瀧澤 雄介

コンピュータ将棋や囲碁等においては，近年のコンピュータ処理能力の大幅な向上により，数手先までのシミュレーションによる盤面評価に基づく着手手の選択が可能となってきている．また，従来の人間による評価関数の設計のみならず，過去のデータやシミュレーションによって生成されたデータを用いた機械学習による評価関数の構築が研究されている．我々は，このような技術進展を踏まえ，これらの技術を 3 次元シミュレーションへ適用することを検討している．本報告では，コンピュータ囲碁において着手手の評価に使用されているモンテカルロ木探索を 3 次元シミュレーションに適用し，着手手の評価に基づく行動選択の効果を確認した．また，非常に計算量の多いモンテカルロ木探索と同等の場面評価をより少ない計算量で実施可能とするため，機械学習を用いた場面評価について検討を行い，実現可能な見通しを得た．

1　はじめに

近年，コンピュータ将棋や囲碁，インベーダー等の 2 次元ゲームにおいて，人間を超える能力を獲得する研究がされている．従来からコンピュータ将棋や囲碁等では，盤面の状態を評価し次の一手を決定するために，棋譜等の過去のデータを元に評価関数を人が作成，調整することで能力向上がされてきた．チェスや将棋等においては，駒の種類や配置等を評価することで人間を超える能力を獲得したが，同一の石を多数使用する囲碁では盤面の評価が難しく，人が作成した評価関数では人間を超える能力を獲得するまでには至らなかった．そのような中，コンピュータ囲碁においてシミュレーションを行い盤面を評価するモンテカルロ木探索 [6] が登場し，従前に比べて飛躍的に能力が向上した．また，近年のコンピュータの計算能力の発展により，脳内の神経回路網を模倣した機械学習手法であるニューラルネットワークを多層化することが可能となり，画像処理において人間の識別能力を上回る [1] とともに，インベーダー等の 2 次元ゲームにおいても人間を超えるスコアを獲得した [3]．この技術はコンピュータ囲碁にも取り入れられ，モンテカルロ木探索と深層ニューラルネットワークを用いた AlphaGo [4] はプロの囲碁棋士に勝利する結果を出している．

このような技術進展を踏まえ，これらの技術を構想検討用の 3 次元シミュレーション [7] のモデルの行動判断に適用することを検討している．本報告では，シミュレーションにより場面を評価するモンテカルロ木探索を 3 次元シミュレーションへ適用した場合の効果の確認結果を示す．また，モンテカルロ木探索は多数のシミュレーションを行い評価を行うものであるため，多数の手を評価するためには，非常に多くの計算資源が必要になる．そこで，機械学習手法を用いて，モンテカルロ木探索を行わずに同等の評価値を算出する手法の検討結果を示す．

2　３次元シミュレーションの問題設定

3 次元シミュレーションとして本報告では，図 1 に示す 1 対 1 の航空機同士の遷移行動を題材とした．問題の簡単化のため，ターン制のゲームと同様に自身の

Toru Ishii, Masanori Yano and Yusuke Takizawa, 防衛装備庁先進技術推進センター, Advanced Defense Technology Center, Acquistion, Technology and Logistics Agency.
Yusuke Chiba and Masahiro Okada, 三菱重工業株式会社, Mitsubishi heavy industries, Ltd.

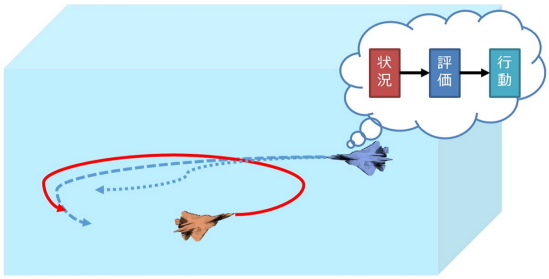

図 1　3 次元シミュレーション概要

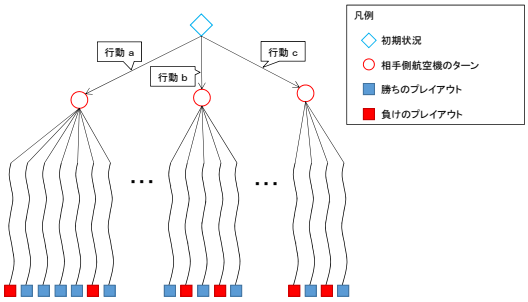

図 2　モンテカルロ木探索

順番においてのみ行動できるものとし，機体の性能及び現在の状態から次の遷移先を制限した上で，相手の後方位置に自身の機体を一定時間保持させることができれば勝利とした．なお，この問題設定においては，機体を質量や長さを持った物理的なモデルとしては計算せず，機体の遷移に条件をかけることにより，物理的に困難な機動を行って勝利するといったことがないように設定した．また，遷移の方法についても無数の選択肢が取り得るため，この問題設定では上昇，下降，右旋回，左旋回及び現状維持といった代表的なもので遷移行動を構成している．

3　モンテカルロ木探索による評価値の算出

モンテカルロ木探索は，モンテカルロ法と木探索を組み合わせたものである．図 2 に示す様に取りえる手をノードとして生成し，各ノードから終局までのシミュレーション（プレイアウト）を行うことでノードの評価値（勝率）を算出するため，評価関数を作ることが難しい問題についても適用できる利点がある．

本報告では，モンテカルロ木探索のアルゴリズムとして UCT(UCB applied to trees) を採用する．UCT は，各ノードに同数のプレイアウトを実施するのではなく，その時点でのプレイアウトの回数と勝率に基づき有望そうなノードに多くのプレイアウトを割り当て，プレイアウトの回数が一定の閾値に達したノードについてはノードを展開し，探索木を成長させることでより効率的に探索するものである．

図 3 に示す 3 次元シミュレーションでの行動判断は，モンテカルロ木探索により勝率及び負率を算出し，もっとも有利な行動を選択するものである．この

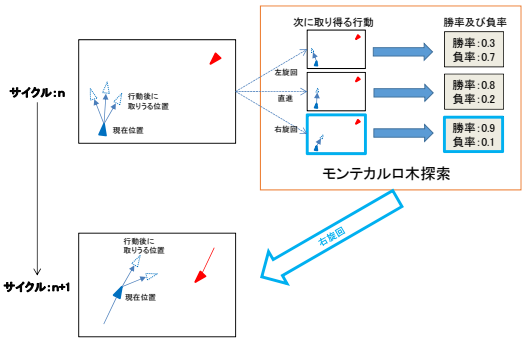

図 3　モンテカルロ木探索による行動判断

行動判断の確からしさを確認するため，シミュレーション開始時の航空機の相対位置を変化させながらシミュレーションを実施した．図 4 から図 6 に航空機の能力を変えてシミュレーションした結果を，相手機を中心とした各位置ごとの勝率として示す．X 軸が機体の左右方向，Y 軸が機軸方向を示しており，Y 軸の正が機体前方である．また，青色が自機の勝率が低い自機の初期位置，赤色が自機の勝率が高い自機の初期位置である．機体性能が同一の場合（図 5）では，相手機の前方から開始した場合では勝率が低く，後方から開始した場合は勝率が高いことがわかる．自機の機体性能を相手機に比べて劣った性能とした場合（図 4）では，機体性能が同一の場合と比較して全体的に勝率が低くなっていることが確認できる．また，自機の機体性能を相手機に比べて優れた性能とした場合（図 6）では，機体性能が同一の場合と比較して勝率が高

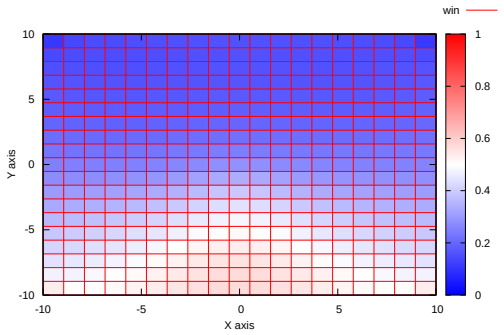

図 4　劣っている性能の場合の勝率

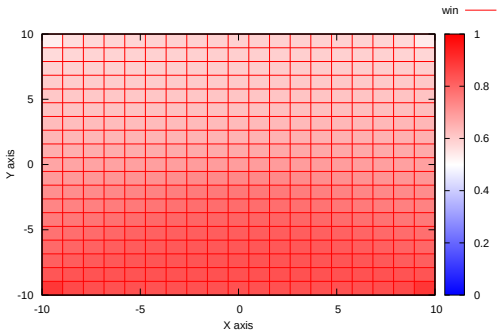

図 6　優れている性能の場合の勝率

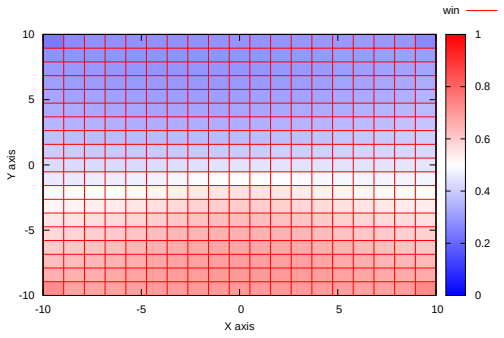

図 5　同一性能の場合の勝率

くなることが確認できる．これらのことは，一般的に想定される機体位置での勝ち負けの感覚と同様であり，モンテカルロ木探索による行動判断によるシミュレーション結果はある程度確からしいと判断できる．

4　機械学習による評価値算出の効率化

モンテカルロ木探索による行動判断の一定の有用性を確認できたが，評価値の算出には複数回のプレイアウトを実施する必要があり，非常に計算コストが高いことが構想検討用のシミュレーションに採用する際の問題となる．我々は，行動判断時に瞬時に場面評価を実施する方法として，畳み込みニューラルネットワークによる画像認識を適用し，場面を画像として認識し評価値を算出することを検討した．畳み込みニューラルネットワークによる画像認識では，学習時は非常に計算コストが高いが，推論時は順方向の計算のみであるため，モンテカルロ木探索のアルゴリズムに比べて計算量を小さくできると考えられる．本報告では，ILSVRC(ImageNet Large Scale Visual Recognition Competition)2014 で最も良い成績を残した畳み込みニューラルネットワーク構造である Google Net のベースとなった，ネットワークインネットワーク [2] を採用した．ネットワークインネットワークは畳み込み層の中に全結合層を持ち，単体の畳み込み層と比較してより鮮明に特徴が抽出された特徴マップの生成が可能となる．また，プーリング層において，最大値プーリングを行うことによって，特徴がより強められた特徴マップを生成する．特徴を強めていくことで，教師データはいくつかのグループにクラスタリングされ，各グループの境界が明確になり識別性能が向上する．

本報告での畳み込みニューラルネットワークの構成を図 7 に示す．3 次元シミュレーションの場面状況を認識させるため，図 8 及び図 9 に示す形式で自機及び相手機の計 4 チャンネルの人工的な画像として生成した．人工的な画像は 3 次元空間における現在の位置座標及び次に取り得る 3 次元空間の位置座標を表現することで，画像認識の問題となるように設定した．本報告では，教師データとして前述の人工的な画像及びそれに対する評価値（勝率及び負率）を 105,666 組作成し，畳み込みニューラルネットワーク

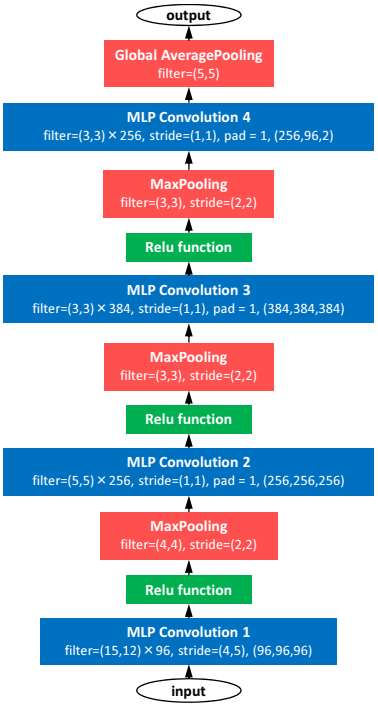

図 7　畳み込みニューラルネットワークの構成

の学習を実施した．なお，畳み込みニューラルネットワークの学習，推論には深層学習フレームワークの Chainer(v1.17.0) [5] を用いた．

畳み込みニューラルネットワークの学習の進捗を確認するため，11,622 組の検証用データを用いて，検証用データにおける勝率及び負率との差分を確認した．図 10 に 100 エポック経過時点までの検証用データによる検証結果を示す．1 エポック経過時点においては，勝率及び負率ともに 10% 程度となっている．70 エポック経過以降の学習が十分に進捗した時点においては，勝率については 3% 台，負率については 2% 台の差分となっており，学習に使用していない未知のデータに対してもモンテカルロ木探索と同等の評価値を出力できていると考えられる．このことから，事前に学習することができれば，行動判断時に計算時間のかかるモンテカルロ木探索を行わずに場面評価ができるものと考えられる．

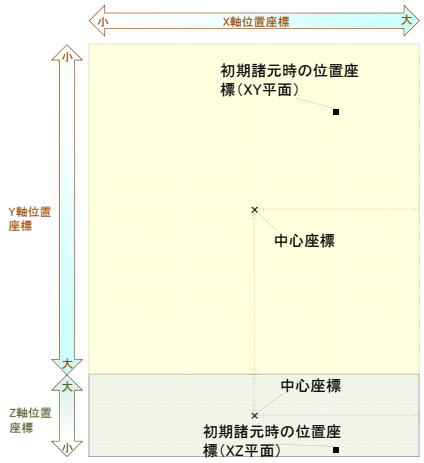

図 8　現在の位置座標に関する画像形式

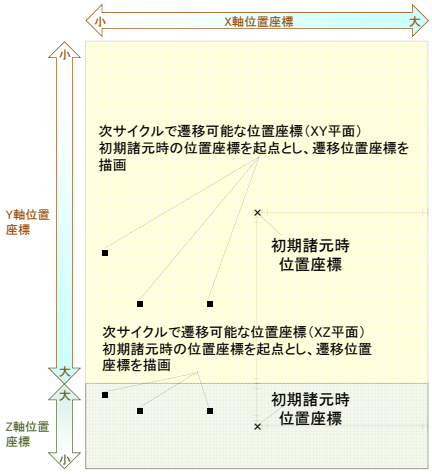

図 9　遷移可能な位置座標に関する画像形式

5　まとめと今後の課題

機械学習技術の発展を踏まえ，これらの技術を構想検討用の 3 次元シミュレーションにおけるモデルの行動判断への適用を検討した．モンテカルロ木探索を用いた行動判断によるシミュレーション結果から，航空機の位置関係から想定される形勢とある程度確からしいことが確認できた．機械学習による評価値算出の効率化は，畳み込みニューラルネットワークを用いることで，学習に使用していない未知のデータに対し

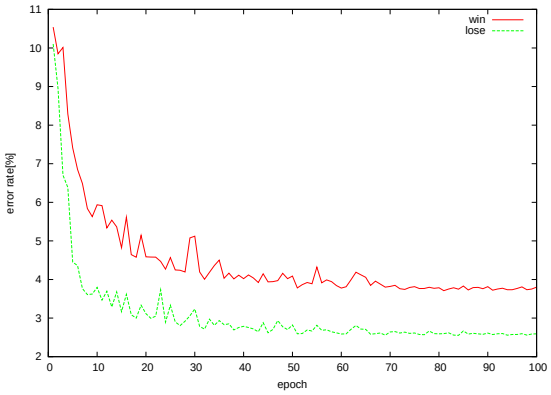

図 10　畳み込みニューラルネットワークによる評価値の学習結果

てもモンテカルロ木探索と同等の評価値を出力できることが確認できた．今後の課題として，多くの航空機等が登場する環境について場面評価を行う方法や，構想検討用の 3 次元シミュレーションに適用する際の要件等を検討する必要がある．

参 考 文 献

[1] He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, *ArXiv e-prints*, (2015).

[2] Lin, M., Chen, Q., and Yan, S.: Network In Network, *CoRR*, Vol. abs/1312.4400(2013).

[3] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning, *Nature*, Vol. 518, No. 7540(2015), pp. 529–533.

[4] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al.: Mastering the game of Go with deep neural networks and tree search, *Nature*, Vol. 529, No. 7587(2016), pp. 484–489.

[5] Tokui, S., Oono, K., Hido, S., and Clayton, J.: Chainer: a Next-Generation Open Source Framework for Deep Learning, *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*, 2015.

[6] 美添一樹: モンテカルロ木探索-コンピュータ囲碁に革命を起こした新手法, 情報処理, Vol. 49, No. 6(2008), pp. 686–693.

[7] 小谷健人, 矢農正紀: ユーザビリティを追求した構想検討用シミュレーションの開発と運用, 第７７回全国大会講演論文集, Vol. 2015, No. 1, mar 2015, pp. 43–44.