F-022

ニューラルネットワークを用いた囲碁のための効率的な強化学習手法
Efficient Reinforcement Learning by using Neural Networks for Go Games

武田 敦志 †
Atsushi Takeda

1. まえがき

2017 年に発表された AlphaGo Zero により、ニューラルネットワークと強化学習を用いることで、過去の棋譜を使わなくてもプロ棋士と同等以上の棋力を有するコンピュータ囲碁のプログラムを実現できることが明らかとなった [1]。一方、AlphaGo Zero の学習処理を実行するためには膨大な量の計算機資源が必要となる。そこで、本稿では、コンピュータ囲碁プログラムのための効率的な強化学習手法を提案する。提案手法では、自己対局における候補手を評価するために領域予測ニューラルネットワークを用いる。この手法を用いることにより、一般的な GPU などの現実的な計算機資源で囲碁プログラムのための強化学習を実行できる。

2. ニューラルネットワークを用いた囲碁プログラム

プロ棋士などの囲碁の実力者が行った対局の棋譜を用いることにより、囲碁の盤面データから次に着手すべき座標を予測するニューラルネットワーク（Policy Network）を実現できる [2]。Policy Network を導入した囲碁プログラムでは、死活などの局所的な状況だけではなく、盤面全体の状況を考慮した候補手を計算できる。また、Policy Network を用いた囲碁プログラムの自己対局から生成された膨大な量の棋譜を用いることにより、盤面データからその対局の勝敗を予測するニューラルネットワーク（Value Network）を実現できる [3]。最新の深層学習技術を用いて実装された Policy Network や Value Network の予測精度は良好であり、これらのニューラルネットワークを活用したコンピュータ囲碁プログラム AlphaGo はプロ棋士と同等以上の棋力を実現している。また、人間が対局した棋譜を必要としないコンピュータ囲碁プログラムの開発が行われている。2017 年に発表された AlphaGo Zero では、過去に行われた対局の棋譜データを使用せず、自己対局による強化学習のみを用いてプロ棋士以上の棋力を達成した [1]。また、Leela Zero・Phoenix Go・ELF OpenGo などの囲碁プログラムでも、AlphaGo Zero と同様の手法を用いることで高い性能を達成している。

一方、AlphaGo Zero の方式を用いてコンピュータ囲碁のプログラムを開発するためには、膨大な量の計算機資源が必要となる。一般的に、高い性能の囲碁プログラムを実現するためには、そのプログラムで用いる Policy Network の性能を強化学習によって改善する必要がある。ただし、Q 学習などのエピソード終了時の報酬に基づいてパラメータを更新する強化学習手法を用いた場合、囲碁の探索空間が非常に大きいため Policy Network のパラメータを効果的に更新できない。そこで、AlphaGo Zero では方策反復法を用いて Policy Network のパラメータを更新している。具体的には、自己対局で着手を計算する際に Policy Network とモンテカルロ木探索を用いて候補手の評価を行い、この評価結果に基づいて Policy Network のパラメータを更新する。しかし、この強化学習手法では候補手を評価するためにシミュレーションを行っており、この計算を現実的な時間で終わらせるためには膨大な量の計算機資源が必要となる。そのため、現実的な計算機資源と計算時間で囲碁プログラムの強化学習を行うためには、より効率的な候補手評価手法が必要となる。

3. 領域予測ネットワークを用いた強化学習手法

本稿では、領域予測ニューラルネットワーク（Territory Network）を用いたコンピュータ囲碁プログラムのための効率的な強化学習手法を提案する。Territory Network とは、囲碁の盤面データから各座標が最終的に黒地・白地・セキとなるかを予測するニューラルネットワークである。提案手法では、AlphaGo Zero と同様に、方策反復法を用いて Policy Network のパラメータを更新する。しかし、AlphaGo Zero が Policy Network とモンテカルロ木探索を用いて自己対局の候補手を評価するのに対し、提案手法では Territory Network のみを用いて候補手を評価する。AlphaGo Zero では候補手を評価するために Policy Network による 1 万回以上の推論を必要とする。一方、提案手法では、候補手を評価するために数回程度の Territory Network の推論のみを必要とする。そのため、提案手法は AlphaGo Zero よりも少ない計算機資源で実行できる。

図 1 に提案手法の自己対局における候補手の評価手順を示す。提案手法では、自己対局において着手座標を計算する場合、最初に Policy Network を用いて候補手の一覧を作成する。ここで、候補手の一覧とは Policy Network が最も大きい着手確率を示した n 個の座標（n は設定パラメータ）である。次に、それぞれの候補座標に着手したときの盤面データを作成し、Territory Network を用いてそれぞれの盤面データの最終的な黒地と白地の大きさの期待値を計算する。ここで、着手をパスした盤面で予測される地の大きさと候補座標に着手した盤面で予測される地の大きさを比較する。このとき、黒番であれば黒地の増加量を候補手の評価値とし、白番であれば白地の増加量を候補手の評価値とする。最後に、最も高い評価値となった候補手の座標に着手する。上記の動作を繰り返し、合理的な着手座

†東北学院大学教養学部情報科学科
Department of Information Science, Tohoku Gakuin Univ.

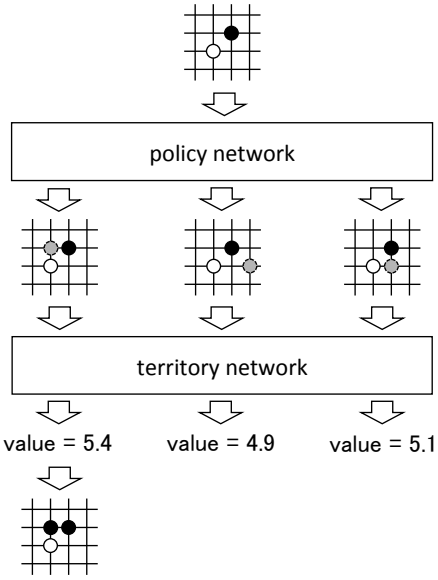

図 1: 自己対局における候補手の評価と着手

標が存在しない状態（セキとなっている箇所以外はパスを続けても黒地・白地が変化しない状態）になると自己対局を終了とする。

複数回の自己対局を行った後、これらの対局の棋譜を用いて Policy Network と Territory Network のパラメータを更新する。まず、棋譜に含まれる勝利した方の着手座標を正解データとして Policy Network のパラメータを更新する。また、棋譜の最終局面を正解データとして Territory Network のパラメータを更新する。自己対局とパラメータ更新を繰り返すことにより、Policy Network と Territory Network の性能を向上させる。

4. 実装・評価

提案手法を評価するため、この強化学習手法を導入したコンピュータ囲碁のプログラムを実装した。この実装では、AlphaGo Zero と同様に、Policy Network と Territory Network を同一のニューラルネットワークとして実装した。ただし、AlphaGo Zero とは異なり、これらのニューラルネットワークの構造は SwGridNet[4]（パラメータ数は 500K）である。このプログラムは自己対局フェーズと学習フェーズを繰り返すように実装した。自己対局フェーズでは 102,400 回の自己対局と棋譜生成を行い、学習フェーズでは自己対局フェーズで生成された最新の 512,000 個の棋譜を用いてニューラルネットワークのパラメータを更新した。

図 2 に自己対局フェーズと学習フェーズを 10 回繰り返したときの Bayes Elo Rating[1]の変化を示す。ここで、Bayes Elo Rating はランダムに着手するプログラムを 0 としたときの相対値となっている。実験結果より、自己対局とパラメータ更新を繰り返すことによりニューラルネットワークの性能が改善し、プログラムの棋力が向上することがわかる。また、一般的な GPU である Nvidia Geforce 1080Ti を用いた場合、48 時間

[1] https://www.remi-coulom.fr/Bayesian-Elo/

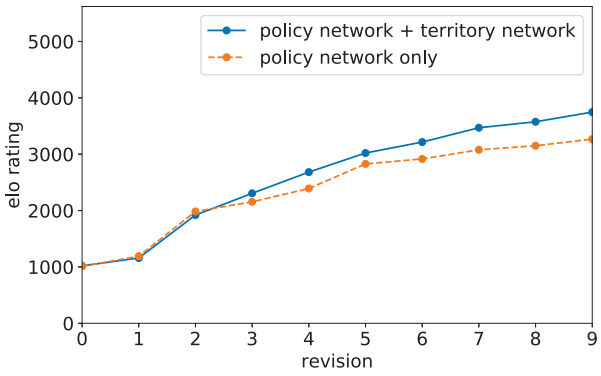

図 2: 強化学習を行ったときの Elo Rating の変化

で 102,400 回の自己対戦を実施できることを確認した。

5. むすび

本稿では、コンピュータ囲碁のプログラムを実現するための効率的な強化学習手法を提案した。提案手法では、自己対局における候補手を評価するために Territory Network を用いることで、既存手法である AlphaGo Zero よりも少ない量の計算機資源で効率的な強化学習を実行できる。提案手法を導入したコンピュータ囲碁のプログラムを実装し、提案手法の有効性を検証した。実験の結果、これらの棋譜を用いた学習によりランダムに着手するプログラムのよりも十分に強い棋力を達成できることを検証した。また、Nvidia Geforce 1080Ti を用いた場合、提案手法は 48 時間で 102,400 回の自己対局と棋譜生成が可能であることを確認した。

参考文献

[1] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

[2] Chris J. Maddison, Aja Huang, Ilya Sutskever, and David Silver. Move evaluation in go using deep convolutional neural networks. In *Proceedings of the Third International Conference on Learning Representations*, 2015.

[3] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[4] Atsushi Takeda. Swgridnet: A deep convolutional neural network based on grid topology for image classification. *arXiv preprint arXiv:1709.07646*, 2017.