WCSC31 詳細アピール文書（PAL）

山口祐

1　開発動機

近年、AlphaGoZero[*1]を発端とする深層強化学習ベースの将棋ソフトは dlshogi[*2]をはじめ開発が加速しており、推論デバイスである GPU の性能向上もあり飛躍的に強化されつつある。

一方で、初期段階から学習手法が一貫した学習の報告は、AlphaZero[*3]の追試を主目的とする AobaZero[*4]のみであり、これとは異なるニューラルネットワークの構造や教師あり学習を組み合わせた場合の学習過程については明らかになっていない点が多い。

そこで PAL（以下、本ソフト）では Squeeze-and-Excitation Networks (SENet) など画像分類タスクで高い性能が報告されているモデル構造やチェス・囲碁ソフト開発における手法などを採用し、深層強化学習を行った。

2　開発過程

ニューラルネットワーク構造として、20 ブロック・256 フィルターの ResNet に Policy/Value 出力用の畳み込み層を持つ AlphaZero 型をベースとした。本ソフトではこれに加えて、ResNet の各畳み込みブロックに SENet ブロック (ボトルネック係数 =8) を挿入し、活性化関数も Relu から Swish[*5]に変更した（表 1)。

また、入力特徴としては先後の駒の位置 (28)、先後の駒の利き (28)、先後の駒の利き数（6)、持ち駒の数（14)、手番 (1) の合計 77 特徴平面の情報を採用した。各数値は持ち駒の数のみその実数、それ以外は 0 または 1 が 9x9 マス分保持される。

Policy の出力特徴は、手番側からみた駒の (移動元, 移動先) のマス（駒打ちの場合は移動先と駒の種類）および成・不成のの組み合わせのうち、合法手として出現しうる 3781 通りをベクトル表現とした softmax 後の方策分布が出力される。また、Value は手番側から見た評価値として [-1.0, 1.0] のスカラー値が出力される。

表1: ニューラルネットワーク構造の比較

	AlphaZero	PAL
入力次元	362 x 81	77 x 81
ブロック数	20	20
フィルター数	256	256
SE ブロック	なし	あり
活性化関数	Relu	Swish
出力次元	1 + 11259	1 + 3781

初期のランダムパラメータからの学習においては、WCSC29 版の PAL (NNUE 評価関数) による評価値および指し手の教師データを別途 5000 万局面生成した。プロ棋士の公式戦棋譜約 15000 からランダムに 32 手目までを選択し、depth12, multipv5 で探索を行った他、探索値の差が最善手から 200 cp 以下の手については一定確率でランダムに選択するように実装し、生成局面の分散化を図った。

得られた上位 5 位までの cp 評価値 q_i に対して

$$v = \frac{1}{1 + \exp(-q_0/k_v)} \tag{1}$$

[*1] https://www.nature.com/articles/nature24270
[*2] https://github.com/TadaoYamaoka/DeepLearningShogi
[*3] https://science.sciencemag.org/content/362/6419/1140
[*4] https://github.com/kobanium/aobazero
[*5] https://arxiv.org/abs/1710.05941v1

$$p_i = \frac{\exp(-q_i/k_p)}{\sum \exp(-q_j/k_p)} \tag{2}$$

ここで $k_v = 600$, $k_p = 100$ とし、初期学習用の状態価値 v, 方策確率 p_i $(i < 5)$ をそれぞれ設定した。また、勝敗価値 z は初期学習では常に v と同じとした。

学習の損失関数は AlphaZero と同様、方策分布の交差エントロピー、価値の二乗誤差、重みの L2 正則化項から構成される。学習用ライブラリには Tensorflow を採用し、最適化アルゴリズムとして Adam を学習率 2.8e-4 から epoch ごとに半減させた。NVIDIA Tesla V100 8 基を用いてバッチサイズ 1024 で 4epoch（2 億局面）分の学習を行い、強化学習の第 0 世代とした。

強化学習では技巧[*6]に囲碁ソフト AQ[*7]の探索部、学習部を移植する形で実装した。1GPU ごとに 256 の異なる対局を割り当て、各対局で 1 手あたり 800 回探索を行うとともに、探索で得られた評価すべき 256 の末端局面をまとめて GPU で推論し、局面生成効率の向上を図った。

強化学習の開始局面はプロ棋士の公式戦棋譜約 15000 の 32 手目までの局面、および floodgate[*8]の 2017 年から 2020 年までの対局のうち、双方レーティング 3000 以上の棋譜で評価値の絶対値が 500 以下、33 手目以降の局面からランダムに選択した。

終局判定については全体の 10%は詰みまたは引き分けまで実施し、90%については 10%の対局の評価値推移を調べ、評価値 x を 5 手連続で下回った場合にその後実際には負けなかった対局が 5%未満になるように、動的に投了閾値を調整した。

局面の生成には平均で Tesla V100 35 基程度を使用した。生成した棋譜は 20000 対局 (約 200 万局面) ごとに 1 世代とし、世代あたりの学習時間は約 2 時間 20 分程度であった。パラメータの更新には Replay buffer として直近 50 世代分の局面データからランダムに 400 万局面を選択した。Value の損失関数としては、局面の探索評価値 v とその対局の結果報酬 z の事情誤差の平均とし、Policy の損失関数は探索開始局面の探索訪問回数を方策分布とみなして交差エントロピーを求めた。0-100 世代目までを学習率 1e-4, 101-150 世代目までを 5e-4、それ以降では 2.5e-5 とし、バッチサイズ 128 で Tesla V100 1 基で 220 世代まで学習を行った。

対局用の探索部については PV-MCTS を採用し、LeelaChessZero[*9]を参考に、ノード構造の省メモリ化、バックアップ完了前の探索回数の管理、複数スレッドによる非同期探索等を実装した。また、詰みルーチンについてはやねうら王[*10]の dfpn 実装を参考にルート局面のみ探索をするようにした。（探索の末端局面については離れ駒以外の簡易 3 手詰判定のみを採用）

探索の推論モデルは TensorRT（7.2.3, CUDA11.1）を用いて GPU ごとに最適化したものを使用した。平手初期局面において GeForce 2080Ti では探索速度が 8600nps だったのに対し、Ampere A100 では 32000nps と約 3.7 倍に向上した（表 2)。また、A100 を 8 基並列に使用した場合でも 230000nps と 90%程度の効率でスケーリングできている。

表2: 平手初期局面の探索速度の比較

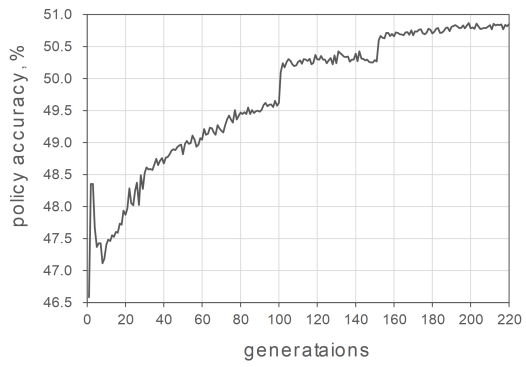
推論 GPU	探索速度 (nps)
GeForce RTX2080Ti	8,600
Ampere A100 x1	32,000
Ampere A100 x8	230,000

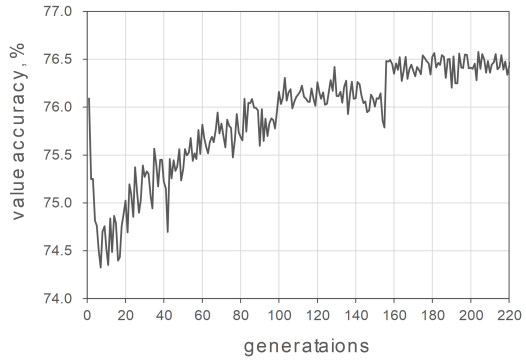
3 結果・考察

学習過程のモデルの評価として、floodgate の 2017 年から 2020 年までのレーティング 3500 以上同士の対局のうちランダムに選択した棋譜から重複を除いた 52 万局面程度を抽出し、テスト用データセットとした。テスト用データセットに対して、各モデルの policy accuracy（一手一致率)、value

[*6] https://github.com/gikou-official/Gikou
[*7] https://github.com/ymgaq/AQ
[*8] http://wdoor.c.u-tokyo.ac.jp/shogi/
[*9] https://github.com/LeelaChessZero/lc0
[*10] https://github.com/yaneurao/YaneuraOu

accuracy (局面評価値の符号と結果の符号の一致率)、value mse（局面評価値と結果の平均二乗誤差）について記録した（図 1)。





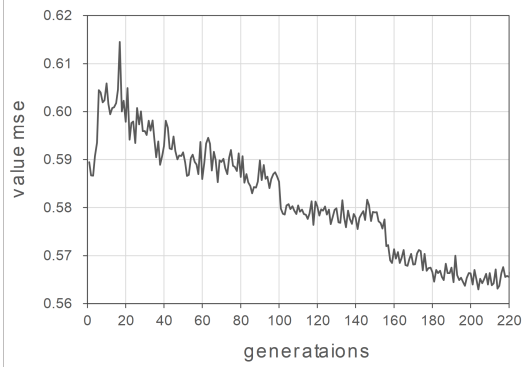


図1: テスト局面の policy/value 一致率の推移

最終的なモデルの評価指標では policy accuracy で 50.8%、value accuracy で 76.5%となり、第 1 回電竜戦バージョンの dlshogi および GCT[*11]の学習モデルをいずれも上回った（表 3)。

表3: テスト用局面の評価指標の比較

	dlshogi	GCT	PAL
policy accuracy	47.2%	47.3%	50.8%
value accuracy	74.9%	75.0%	76.5%
value mse	0.587	0.600	0.565

本ソフトの強化学習に使用した計算資源は Tesla V100 換算で延べ 20000GPU 時間程度であり、GeForce RTX3090 など一般向けの GPU であっても 4 基で 7 ヶ月程度あれば十分再現できると考えられる。

[*11] https://gist.github.com/lvisdd/9b49ab88600fa242f2138fad4eb06caf