AobaZero のアピール文書

山下 宏
yss@bd.mbn.or.jp

1 AlphaZero の追試が最初の目的

AobaZero は Bonanza、LeelaZero のコードをベースに AlphaZero の追試をするべく MCTS ＋ディープラーニングで実装されてます。ネットワークは 3x3 のフィルタが 256 個の 20 block の ResNet でパラメータの個数は 2340 万個。棋譜生成をユーザの皆様と協力して行う分散強化学習です。オープンソースです[*1]。

2 AlphaZero の追試は 2021 年 4 月に終了

AlphaZero の将棋の追試は、2019 年 3 月から開始し、2021 年 4 月に 3900 万棋譜を作成して終了しました[*2]。その後、表 1 の変更を行っています。2022 年 3 月 30 日現在、5360 万棋譜を作成しています。

表 1　追試終了後の改良

日付	
2021 年 4 月	40 block に移行
2021 年 9 月	20 block に戻し温度を 1.0 から 1.3 に変えて序盤の変化を増やす
2021 年 12 月	Value の学習を「実際の勝敗」から「実際の勝敗」＋「探索勝率」の平均に
2022 年 1 月	1 手 800playout 固定から 100〜3200 と可変に
2022 年 2 月	ネットワークの構造を dlshogi 風に変更 利きの情報あり、ReLU を Swish に 30 手目までのランダム性を変更

3 40 block に移行

まずネットワークのサイズを 20 block から倍の 40 block に変更しました。囲碁の KataGo では +150 Elo ほどの効果があり、この程度の上昇を期待していたのですが実際は +40 Elo 程度でした。原因はいくつか考えられます。

- 学習が収束した状態、の棋譜から再学習したもので開始したせい。最初から 40 block だと違う？
- 学習が収束した状態で、30 手後のユニークな局面は 40% 程度 (100 万棋譜で)。同じような (相掛かりの) 棋譜ばかりで多様性がない。30 手後の細かい定跡の変化で強くなっている？
- そもそも将棋では 20 block 以上にしても効果が少ない？ 20 block で作った棋譜を学習させた 10 block は -120 ELO 弱い。
  40 block +40 ELO 強い
  20 block
  10 block -120 ELO 弱い
- 40 block の再学習で学習率の下げ方が早すぎた (Cosine Annealing を 1 回だけ、の方が良かった？)。
- バグ

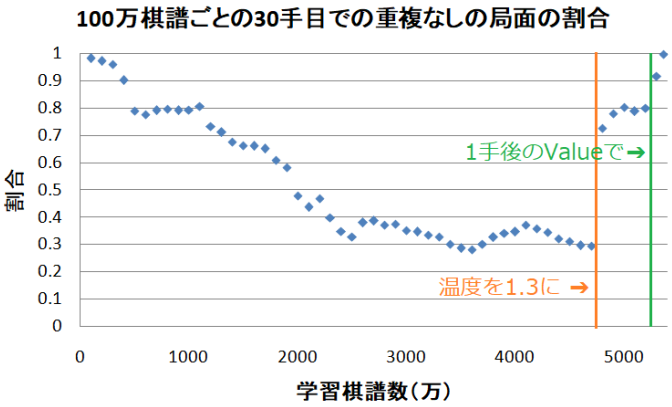

図 1　重複なしの (ユニークな) 局面の割合

4 重複局面を減らすために温度を 1.3 に

図 1 は学習棋譜の 30 手目で重複していない (ユニークな) 局面の割合です (100 万棋譜ごと)。学習開始時はほぼすべての棋譜がバラバラですが、徐々に同じ棋譜を生成するようになり、AlphaZero が学習を終了した 2400 万棋譜では 35% 程度まで下がります。具体的には初手から▲ 26 歩△ 84 歩の相掛かりの将棋ばかり指すようになります。重複を減らそうと 4700 万棋譜の時点で温度[*3]を 1.0 から 1.3 に変更して、初手から 30 手までは訪問回数が少ない手も選ばれやすいよ

[*1] https://github.com/kobanium/aobazero
[*2] https://github.com/kobanium/aobazero/issues/54
[*3] 温度→ 0 で訪問回数最大の手を、温度 1 で訪問回数の割合で、温度→∞ですべての手を均等に選ぶ

うにしました。これで重複なしは 80% まで上がります。が、実際は途中で悪手を多く指してるため形勢に差がつくことが多く 4 割近い棋譜が 31 手目で投了していました。この変更による棋力向上はわずかです (+33 ELO)。

5 「実際の勝敗」＋「探索勝率」の平均を学習

その後、Value の学習を「実際の勝敗」から「実際の勝敗」＋「探索勝率」の平均、に変更しました。棋力には変化なしです。

6 800playout 固定でなく 100〜3200 まで可変に

これはチェスの Leela Chess Zero(LC0) で使われてる手法で[*4] 、1 手 800playout 固定でなく、100〜3200 まで可変にします。生成される棋譜の棋力は +76 ELO 強くなってます[*5]。これは 100playout した時の訪問回数の分布と 200playout での分布を比べて分布に変化がないなら打ち切る、という手法です。カルバックライブラー情報量を使って判定します。LC0 の記述だと kldgain=0.0000013 です。AlphaZero 方式では Policy の分布を学習するので、悪影響は少ないかもしれません。ただこの変更でも、ニューラルネットワーク (NN) の重みの強さは変化なしでした。

7 ネットワークの構造を利きあり、Swish に

AobaZero には長い利きをうっかりする、という欠点があります。例えば図 2 で△ 91 馬 (19) と馬をただで取る手が指せません。この手の着手確率は極めて低く、141 個ある可能手の 141 番目です[*6]。原因は 3x3 のフィルタをたくさん並べた CNN の構造では距離が遠い位置関係の認識が苦手なためです。囲碁でもシチョウ、と呼ばれる同じく距離が長い一本道の探索の認識が不得意です。これは NN の着手出力を dlshogi と同じ、あるマス目に 8 方向のどこから移動してきたか[*7]、に変えることでかなり改善され、さらに NN の入力に利きの情報を入れることでほぼ理解できます。着手出力を dlshogi と同じにし、利きの情報も入力とした Aoba 駒落ち[*8] では図 2 の手は 1 番目の候補になり簡単に指せます。

また活性化関数も dlshogi や PAL を参考に ReLU から Swish に変更しました。

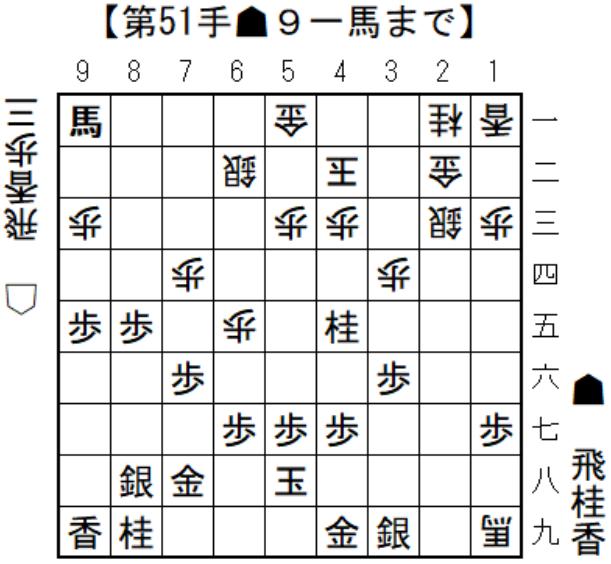

図 2 △ 91 馬 (19) と取る手を読み抜け

8 ランダム性は探索なしの Policy で。互角に近い局面のみを

NN の構造の変更と同時に、多様な棋譜を生成する仕組みも変更しました。今までは 800playout した後、訪問回数の分布から乱数で選ぶ、としていたのを単に Policy の確率で選ぶ、にしました。今まではノイズで変な手を試しても評価値が低いため、訪問回数は増えず選ばれにくかったです。また 1 手指した後の勝率 (Value) が 0.41 < (0.61) < 0.81 の間に収まらない場合は、1 手戻して、直前の手は Policy の最善手を選ぶようにしています[*9]。そして 30 手後の Value が上の範囲を超えた場合は、取り消して最初からやり直します。それと初手から 10 手後程度までは Polciy でなく、1 手指した後の Value の値[*10] を元にしています。これは初期局面で

- ▲ 26 歩の確率が 35.2%、直後の Value が勝率 55.9%
- ▲ 76 歩の確率が 0.6%、直後の Value が勝率 54.5%

と Value は大差ないのに▲ 76 歩の確率が低いのを防ぐためです。最善手以外の Policy は時々極端な値が付くことがあるようです。

[*4] https://medium.com/@veedrac/leela-chess-test40-test50-and-beyond-c15896becfac

[*5] 平均 777playout/手、ほぼ思考時間は同一

[*6] 利きありでの学習後、1 番目になりました (w3924)

[*7] AlphaZero では移動元から 8 方向に何マス移動したか、で 11259 通り。dlshogi は 2187 通り。

[*8] 駒落ちをゼロから深層強化学習させたもの。http://www.yss-aya.com/komaochi/

[*9] 0.61 は過去 100 万棋譜での先手勝率

[*10] 最善の Value との差を diff とすると、1/exp(diff*70)

30 手までの手順が決まれば、少なくともその手を 1 回は探索するようにして、800playout 後に強制的にその手を選んでます。

これらによって棋譜の重複なしの割合は 99.1% とほぼ完全にバラバラになってます。他にすべてのノードで 3 手詰を調べるようにしました。これは +20 ELO 程度の向上でした。そもそも 1 手詰の形や 3 手詰の形は簡単なようで、NN が「覚えて」しまっています。これらの変更でも現状、棋力に変化はなく、正しい方向なのかは不明です。

9　人間の知識は使っていない、をおそらく継続

利きの情報の追加や 3 手詰などで AlphaZero からは離れてきましたが、まだ全体としては「人間の知識は使っていない」を継続していると考えています。

10　1 手 1playout で将棋クエストで 6 段に

AobaZero の w3880[*11]が 1playout で将棋クエストの :FuriJirouBot というアカウントで長考 (持ち時間 10 分) で 6 段 (2250 点) になっています。名前から推測できるように振飛車しか指さない設定です[*12]。序盤はユーザ棋譜から作った定跡で振飛車を選ぶようになってるそうです。現在の AobaZero は振飛車を指さないのですが (初期の 400 万棋譜時点では後手のみ四間飛車を指していました) それでも学習棋譜では時々出てくるのでそれなりに指せるようです。

floodgate で 1 手 1playout は 2150 点 (w3392) でした。将棋クエストの長考、のレートはほぼ一致するようです。floodgate は対戦相手が偏っているのと勝率が低い (9 勝 83 敗、勝率 0.10)、アンカー (3300 点) から離れてる、でやや不確かですが。

余談ですが囲碁の KataGo も 1playout で囲碁クエストで :Katago1pBot で動いており下の段位になっています。

- 9 路 で 2470 点 (7 段)
- 13 路で 2820 点 (9 段)
- 19 路で 2750 点 (9 段)

11　NN は内部で探索してる？

まったく探索なしで局面を NN に与えて返ってくる最善手を指すだけでこれだけの強さがあるのは驚きです。ただ探索してないとはいえ、NN の内部ではおそらく複雑な if 文の組み合わせで疑似的な探索をしている (例えば、この形は角を切って同玉なら頭金で詰む形なので角を切る手の確率が高い、みたいな) ので、まったく読んでいない、というのはやや語弊があるのかもしれません。

12　詰を読むと水平線効果の無駄な王手をするように

また棚瀬さんから指摘されたのですが、3 手詰を読むようにして学習させた NN の重みは 1 手 1playout でも負けの局面で無意味な王手をするようになりました。これは探索部が負けの局面では王手以外の手を指さないようになり、NN もそれを学習したためです。きれいな形作りには投了直前は人間の棋譜からのみ学習させる、などが必要なのかもしれません。

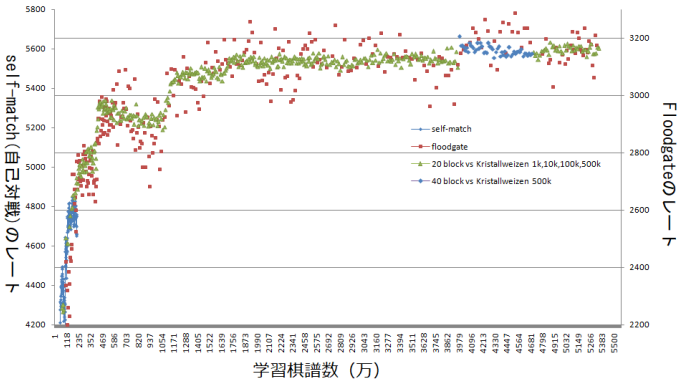

図 3　AobaZero の棋力の推移。右軸が floodgate レート

13　で、強くなったの？

去年からほとんど強くなっていません。強くするのは大変ですね・・・。図 3 が ELO の推移です。

14　3 年で 5300 万棋譜

5300 万棋譜、という膨大な棋譜を 3 年間で生成してきました。棋譜生成に協力していただいてる皆様に感謝いたします。

[*11] w3880 で 3880 番目に作成された重み、を意味します。w3880 は利き情報を使ってない最後の重みでもあります。

[*12] 将棋クエストはトライルール、AobaZero は宣言なので入玉になりにくいように振飛車に、とのことです。