将棋81万：強化学習のための多様性を持った将棋初期局面集

出村 洋介[1,a)]　金子 知適[1,b)]

概要：経験の多様性と不偏性は強化学習エージェントの性能や頑健性を向上させるが，大きな計算コストなしにそれを実現するのは困難な場合がある．多くのチェスライクゲームやオセロなどでは，初期状態（初期局面の駒配置等）が固定されていて 1 通りしかないため，AlphaZero スタイルの強化学習を行う場合，エージェントは似たようなエピソードや棋譜を経験しがちである．本論文では，この課題に対応するため，将棋の初期局面を拡張した「将棋 81 万」を提案し，将棋における有効性を実験的に評価する．「将棋 81 万」は，チェス 960 [1] と同様に駒の初期配置を一定の制約のもとでランダムにシャッフルして作成された将棋の初期局面集である．我々は，Gumbel AlphaZero の手法で 1000 万局の自己対局を行って様々なエージェントを訓練する実験を行い，最初に将棋 81 万で事前学習を行った後に通常の将棋に適応学習させたエージェントは，通常の将棋のみで訓練したエージェントよりも人間の対局で見られる様々な戦型において平均的パフォーマンスや頑健性が向上することを示した．

キーワード：ボードゲーム，将棋，チェス 960，強化学習

Shogi816K: A Diverse Collection of Starting Positions for Reinforcement Learning in Shogi

Yosuke Demura[1,a)]　Tomoyuki Kaneko[1,b)]

Abstract: While the diversity and unbiasedness in experiences will improve the performance and robustness of reinforcement learning agents, it is sometimes difficult to realize them without incurring significant costs. Many chess variants and Othello are typical domains where agents experience similar episodes (or game records) in AlphaZero-style reinforcement learning because there is a single fixed opening state that restricts the legal moves. In this paper, we address the problem by carefully augmenting opening positions to propose Shogi816K and empirically evaluate the effectiveness in shogi, a Japanese chess variant. As in Chess 960 or Fischer Random Chess [1], Shogi816K randomizes pieces in the opening positions with reasonable restrictions. We trained various agents by Gumbel AlphaZero with ten million game records and showed that agents first pre-trained with Shogi816K and later adapted to the usual shogi achieved better performance in average and robustness with respect to various opening variations in human playing than those trained only with the usual shogi.

Keywords: board game, Shogi, Chess 960, reinforcement learning

1. はじめに

エージェントが環境との相互作用を通じて学習を行う強化学習は，囲碁 [2]，チェスや将棋 [3]，ビデオゲーム [4]，ロボット制御 [5] など，様々な分野での応用が進んでいる．

その強化学習の分野で，近年，エージェントの経験の多様性がエージェントのパフォーマンスに与える影響が注目されている．例えば，ビデオゲームの強化学習では，手続的生成 (procedural generation) によりステージのレイアウトや敵の配置などにランダム性を加えた環境で学習させた方が，過学習が生じにくくエージェントがより頑健な方策が獲得できるとされる [6]．また，ロボット制御の分野では，シミュレーション環境で強化学習したロボットを現実世界で動かす場合，シミュレーション環境では物理法則の再現度やセンサー類のノイズなどの点で現実世界とギャップがあるため，シミュレーション環境のカメラ位置や背景色な

[1] 東京大学大学院総合文化研究科
Graduate School of Arts and Sciences, The University of Tokyo
[a)] demura@g.ecc.u-tokyo.ac.jp
[b)] kaneko@graco.c.u-tokyo.ac.jp

どに様々なランダム性を加えた方が，現実世界とギャップが緩和され汎化性能や頑健性が高まるとされる [7].

こうした研究から着想を受けて，本研究では，人工知能研究の対象として長年親しまれている将棋を題材に，多様性を持った環境で学習させた方が AI エージェントが様々な将棋の戦法を上手く指せるようになるか検証を試みる.

将棋は，実現可能局面数が 10^{68}〜10^{69} 程度と推定されている複雑なゲームである [8]．一方で，将棋の駒の初期配置は 1 通りであり，局面の多様性は初手からの分岐としてのみ表れるため，序盤では重複する局面をなぞることも多い．そこで，我々は，序盤から積極的にエージェントに多様な経験を積ませるため，チェス 960 [1] と同様に一定の制約のもと駒の初期配置をシャッフルした初期局面集「将棋 81 万」を考案した (図 1).「将棋 81 万」には左右対称の局面を除き 81 万 6480 通りの初期局面が含まれている.

本論文では，「将棋 81 万」の多様な初期局面を用いてエージェントを強化学習することで，様々な将棋の戦法をうまく指しこなせる AI プレイヤの作成を目指した.

我々の考える本研究の貢献は以下のとおりである.

- 多様性を持った将棋の初期局面集「将棋 81 万」を強化学習の環境として用いることで，学習アルゴリズムは一切変更せずに，振り飛車を含む多様な戦型をより上手く指す AI プレイヤが作成できることを示した.
- ボードゲームの一種である将棋においても，経験の多様性がエージェントの学習を助ける可能性を示した.

図 1 「将棋 81 万」の初期局面の例．通常の初期局面から歩以外の駒の位置をシャッフルすることで作成される.

2. 関連研究

2.1 ボードゲームにおける強化学習

将棋・チェス・囲碁などのボードゲームでは，盤上の初期状態が 1 種類しかなく，途中でサイコロなどのランダム要素も入らない．このため，こうしたボードゲームでは，プレイヤの指し手選択が決定的だと最初から最後の手まで毎回同じ展開になってしまう問題がある.

その対策として，AlphaZero [3] では，エージェントの方策に乱数を加えることで，自己対局中の指し手選択にランダム性を加えて，特定の展開の棋譜ばかりにならないような工夫がされている．また，エージェントの指し手選択を多様化する手法として，損失関数に方策のエントロピー最大化項を加えるエントロピー最大化強化学習が提案されている [9].

加えて，多様性を持つ複数のエージェントのチームを構成し，全体としてエージェントを多様化するという方向性も提案されている [10].

これらの先行研究と本研究は，強化学習時のエピソードに多様性を持たせるという目的は同じだが，その目的達成の手段が異なるという関係にある．先行研究ではエピソード多様化の手段としてエージェントの振る舞いの多様化を図っているのに対し，本研究ではエージェントや学習方法には手を加えずに，局面遷移のルールを維持しつつエピソードの多様性を増加させることを試みている．そのため，先行研究と本研究は，相互に補い合う関係にある.

2.2 データ拡張

画像認識の教師あり学習の研究では，データ拡張（data augmentation）が有力な手法として使われている [11]．例えば，教師あり学習を行う際に，画像を左右反転させる，一部を隠す，他の画像と混ぜるといった加工をデータに加えて学習すると，モデルの予測精度やロバスト性を向上させる効果があることが知られている [12].

強化学習の領域でも，データ拡張が注目されつつある．例えば，DeepMind Control Suite，ProcGen，OpenAI Gym などのピクセル画像を入力とする強化学習において，入力画像をそのまま学習するよりも，入力画像にデータ拡張を行いながら学習する方がデータ効率やパフォーマンスが向上することが示されている [13].

また，ロボット制御の研究では，Domain Randomization [7] という手法が提案されている．ロボット制御の場合には，シミュレーション環境では物理法則の再現度やセンサー類のノイズなどの点で現実世界との間にギャップがあるため，シミュレーション環境下で学習したロボットが実世界環境で上手く動くとは限らない．そこで，訓練時のシミュレーション環境に様々なランダム性を加えることで，実世界環境における頑健性が上がるとされる.

これらの研究は，学習アルゴリズムに変更を加えることなく，学習対象となるデータ側に変更を加えることでエージェントの学習を助ける手法である．本研究もこれらの研究の着想に影響を受けている.

2.3 手続型生成

ビデオゲームライクな環境における強化学習の研究で

は，主にエージェントの汎化性能をテストする目的で，手続型生成（procedural generation）を用いて生成された多様性に富む環境が提案されている．

例えば，ビデオゲーム環境の ProcGen [6] は，手続型生成により，ステージのレイアウトや，ゲーム上の物体の配置，敵の位置や出現間隔などにランダム性が加えられた環境である．ProcGen のような環境では，単にステージを暗記するような方法では目標を達成できないため，エージェントは様々な状況に対して頑健なポリシーを学ぶようになるとされる．

主に訓練用途か評価用途かという違いはあるが，本研究の「将棋 81 万」は，手続的生成類似の手法をボードゲームの初期局面に適用したと考えることもできる．

2.4 チェス 960

チェス 960 (Chess960．別名 Fischer random chess) [1] は，チェスの世界チャンピオン Bobby Fischer によって発案されたチェスの変則ルールの 1 つである．

通常のチェスでは駒の初期配置は固定だが，チェス 960 では駒の初期配置が特定の制約のもとでランダムにシャッフルされるため，合計 960 通りの駒の初期配置が存在する．初期局面において駒の並び替えが行われるほかは，キャスリングなどの一部の例外を除き，チェス 960 のルールは通常のチェスと同様である．

チェス 960 は，人間が序盤定跡の暗記に頼らず創造性を発揮して楽しむ趣旨で考案された．チェスプログラムにもチェス 960 の対局機能を備えたものがあるが（Stockfish[*1], Leela Chess Zero[*2]など），AI エージェントをチェス 960 で強化学習すると通常のチェスが上手くなるか否かに関しては，我々の知る限りこれまで詳細な研究は発表されていない．

「将棋 81 万」は，チェス 960 と同様に駒の初期配置をシャッフルしており，「将棋 81 万」はいわばチェス 960 の将棋版といえる．

3. 提案手法

3.1 「将棋 81 万」(Shogi816K)

AI エージェントにも積極的に多様な将棋を経験させた方が，様々な戦法で将棋がより上手くなるのではないか，という仮説を本稿では提示する．

その検証のために，我々は，強化学習エージェントに多様な経験を積ませるための環境として，「**将棋 81 万**」(**Shogi816K**) という初期局面集を用いることを提案する (図 1)．

将棋 81 万の初期局面集は，以下の制約のもとで駒の初期配置をランダムにシャッフルして作成する．

(1) 先手の歩は通常の将棋と同様に七段目に配置
(2) 先手の飛角は八段目にランダムに配置
(3) 先手の玉金銀桂香は九段目にランダムに配置
(4) 後手の駒は先手の駒と回転対称に配置

将棋 81 万とチェス 960 の制約の違いは，将棋とチェスのルールの違い（駒配置が線対称か回転対称か，キャスリングと呼ばれる特別な手の有無，ビショップの個数と角の枚数の違いなど）を踏まえたものである．

なお，チェス 960 ではキャスリング等の一部ルールを通常のチェスから変更しているが，将棋 81 万ではそのようなルール上の例外はなく，駒の初期配置をシャッフルする以外は通常の将棋のルールと完全に同一である．

以上の条件をみたす駒の初期配置は，左右反転した局面を同一とみなすと 81 万 6480 通りであるため[*3]，上記の制約のもとで作成された初期局面集を「将棋 81 万」と呼ぶこととした．

3.2 事前学習と適応学習

画像認識の分野では，ImageNet などのデータセットで事前学習を行ったのち，類似した別のタスク（物体認識タスクや人体のポーズ推定タスクなど）でファインチューニングを行う手法がよく用いられている．事前学習を取り入れることで，特定のタスクのみで学習を行なった場合よりもモデルの性能や頑健性が向上するといった効果があるとされる [12]．

この方針に倣い，まず将棋 81 万の環境下で事前学習を行い，その後に通常初期局面に環境を変更してエージェントに序盤定跡を学習させる方法も試すこととした．

4. 実験

4.1 学習環境

将棋 81 万がもたらす多様性の効果を検証するため，本研究では，以下の 5 種類の環境下で，自己対局により方策 (policy) と価値 (value) 関数のネットワークを並行して学習させる AlphaZero スタイルの強化学習によってエージェントを訓練した．

- **Base**: 100%通常初期局面を用いた自己対局のみで学習．初期局面は固定のため 1 通り．
- **Shogi816K**: 100%将棋 81 万（Shogi816K）を用いた自己対局のみで学習．初期局面は 816,480 通りから一様ランダム選択．
- **Shogi816K+Base**: 最初の 60%は将棋 81 万で事前学習を行い，その後の 40%は通常初期局面の自己対局で学習．なお，60%という事前学習の割合は実験的に

[*1] https://stockfishchess.org
[*2] https://lczero.org

[*3] 飛・角の配置が $9 \cdot 8 = 72$ 通り，玉・金・銀・桂・香の配置が $9 \cdot {}_8C_2 \cdot {}_6C_2 \cdot {}_4C_2 \cdot {}_2C_2 = 22,680$ 通りあるため，合計で $72 \cdot 22680 = 1,632,960$ 通りの駒の配置が存在する．しかし，将棋のルールにはチェスのキャスリングのような左右非対称のルールがないため，左右反転を同じとみなすと 816,480 通りとなる．

決定した．

- **Shogi816K+Random2ply**: 最初の 60%は将棋 81 万で事前学習を行い，その後の 40%は下記の Random2ply の初期局面で学習を行う．
- **Random2ply**: 初期局面から先手・後手 1 手ずつ合計 2 手一様ランダムに進めた局面を初期局面として 100%自己対局を行う．2 手ランダムに進めた初期局面は 900 通り．将棋 81 万との比較用として検証．

各環境を学習コストの観点で公平に比較するため，上記いずれの環境でも自己対局数は事前学習込みで合計 1000 万対局に統一した．

また，比較対象として，通常初期局面からランダムに 2 手進めた局面を初期局面に用いた環境（Random2ply）を用意した．2 手にした理由は，奇数手だと先手・後手の非対称性が生じること，4 手以上にすると局面数は増やせるが初手から激しい手順（王手や実質角損の手順など）が出現してしまうことを考慮したためである．

4.2 エージェントの訓練

本研究のフォーカスは学習時の環境同士の比較を行うことにある．このため，上記 5 種類のどの環境下で学習する際も，強化学習のアルゴリズムには同じものを利用した．

具体的には，強化学習アルゴリズムには，AlphaZero [3] の後継として発表された Gumbel AlphaZero [14] を用いた．Gumbel AlphaZero は，より少ない学習資源でも強化学習を可能にした AlphaZero の改良版であるが，局面の状態価値（value）と指し手選択の方策（policy）のニューラルネットワークを自己対局を通じて強化学習する枠組みは共通である．

今回の実験では，損失関数 $\mathcal{L}$ は式 (1) を用いた．value については AlphaZero 同様に二乗誤差損失を，policy については実装が簡単な simple policy loss (文献 [14] の Section 4 参照) を用いている．

$$\mathcal{L} = \underbrace{(z-v)^2}_{\text{value loss}} - \underbrace{\log \pi(a)}_{\text{policy loss}} + c||\theta|| \qquad (1)$$

ここで，z は対局結果 (+1: 勝ち, 0: 引き分け, −1: 負け)，v はネットワークの出力する value，π はネットワークの出力する policy，a は自己対局中に実際にエージェントが選んだ次の 1 手，c は L2 正則化係数である．

強化学習の自己対局では，計算資源節約のため，Planning with Gumbel (文献 [14] の Algorithm 1 参照) を用いて，1 局面あたり 16 シミュレーションで対局を行なった．AlphaZero では 1 局面あたり 800 シミュレーションであるため，この設定だと 1/50 のノード数となるが，この条件で将棋でも一定の棋力が得られることは確認済みである (付録 A.2 参照)．その他の強化学習時の設定は付録 A.1 に記載した．

4.3 エージェントの棋力評価

上記の手法で強化学習を行なったエージェントの棋力評価は，実績のある将棋プログラム「水匠 5」[*4] と対戦させて評価した．水匠との対局条件は，以下のとおりである．

- 今回作成したプレイヤの設定：自己対局と同じ条件を用いて，Planning with Gumbel を用いて 1 手 16 シミュレーションで指し手を選択する．並列探索や定跡は利用しない．
- 水匠の設定：探索局面数は 1 手 1000 ノードとし，置換表サイズは 16MB に設定した．並列探索や定跡は利用しない．
- 対局条件：多様な戦法における棋力を評価するため，人間の棋譜集[*5] を用いて 9 つの戦型（角換わり，相掛かり，横歩取り，矢倉，中飛車，四間飛車，三間飛車，向飛車，相振飛車）の途中局面から対局を行なった（20 手目から 40 手目までランダムに選択）．
- 対局数：上記 9 戦型について，各 2000 局実施した．

今回の評価方法では，今回作成したプレイヤ側も水匠側も最高パフォーマンスの設定にはなっていない．本研究は，環境と環境の比較をすることにフォーカスを当てており，他プログラムと強さの比較が主眼ではないため，計算資源の観点も考慮してこのような制限を用いた．

5. 実験結果

5.1 自己対局時の環境がプレイヤの棋力に与える影響

まず，学習環境の多様性の違いが棋力に与える影響を調べるため，作成した 5 種類のプレイヤを水匠と対局させ，水匠との Elo レート差を測定した．各プレイヤの対局結果の概要は図 2 の通りである．

将棋 81 万のみで訓練した場合（Shogi816K），通常初期局面での訓練（Base）と比較して，振り飛車や相振飛車の戦法では棋力向上が見られたが，居飛車の将棋に弱くなる傾向が見られた．将棋 81 万では飛車の初期配置が 1 筋から 9 筋までランダムに配置されるため，これが振り飛車の学習には貢献した一方で，居飛車の学習は進みにくかった可能性がある．

これに対し，将棋 81 万を事前学習として用いる提案手法（Shogi816K+Base, Shogi816K+Random2ply）では，いずれの戦法でも棋力が向上していることが分かる．また，居飛車での棋力向上は Elo レーティング 30 点程度であるが，振り飛車・相振飛車での棋力向上は+50 Elo 程度となっており，Base で比較的苦手だった振り飛車系の強さが底上げされている．

また，通常初期局面から 2 手ランダムに指す手法（Ran-

[*4] https://github.com/mizar/YaneuraOu/releases/

[*5] https://web.archive.org/web/20190330074233/http://www.geocities.jp/shogi_depot/．なお，この棋譜集には，上記 9 戦型はいずれも 2000 棋譜以上含まれている．

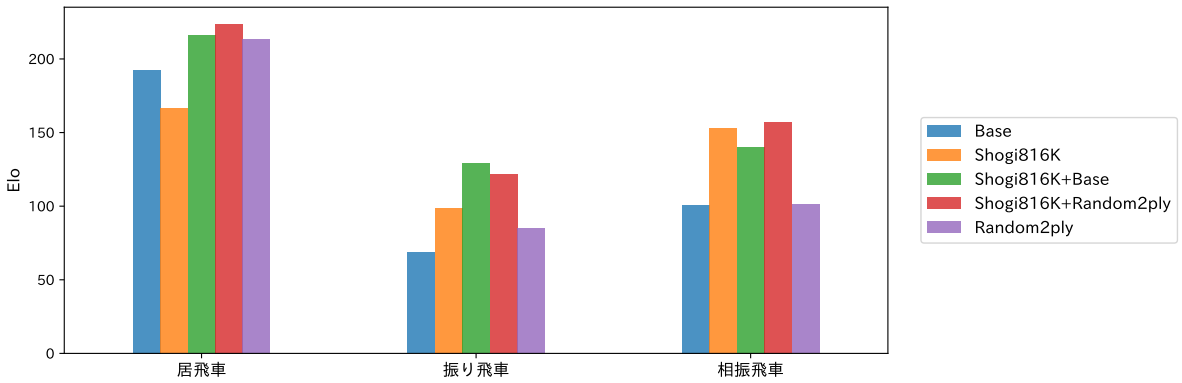

図 2 自己対局時の環境がプレイヤの棋力に与える影響．水匠 5 (1,000 ノード) を ± 0 Elo として Elo レートを測定．居飛車は角換わり・相掛かり・横歩取り・矢倉の平均，振り飛車は中飛車・四間飛車・三間飛車・向飛車の平均．

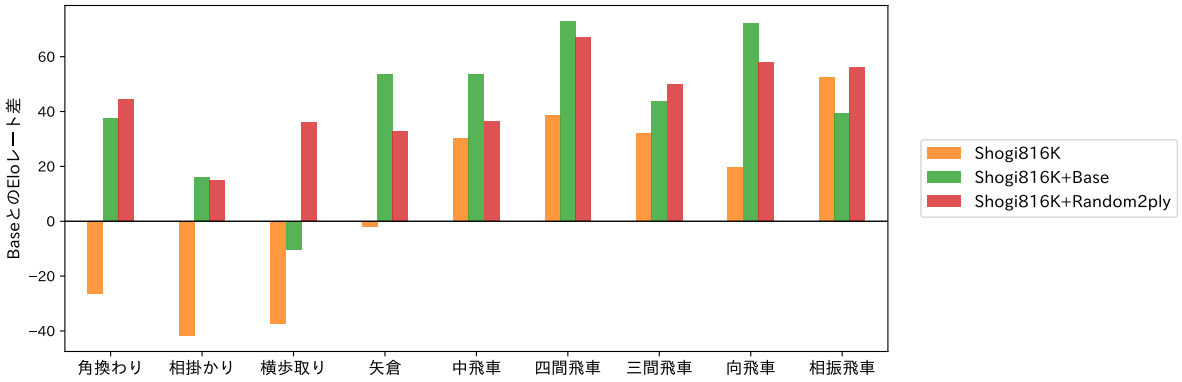

図 3 将棋 81 万の事前学習の効果．図 2 と同じく水匠 5 (1,000 ノード) との対局で各プレイヤの Elo レートを計測した後，通常初期局面のみを学習したプレイヤ (Base) との Elo レート差を戦型別に表示した．

dom2ply）では，振り飛車ではあまり効果がなかったが，居飛車での棋力向上が見られた．Random2ply の多様性は，将棋 81 万ほどではないものの，ある程度エージェントの棋力向上に役立っていることが分かる．

5.2 将棋 81 万による事前学習の効果

将棋 81 万の環境の影響をより詳細に確認するため，通常初期局面のみで訓練したプレイヤ（Base）との Elo レート差を 9 種類の戦型別に表示したものが図 3 である．図 3 では，図 2 と同じく水匠との対局により各プレイヤの Elo レートを測定した後，各プレイヤと Base とのレーティングの差を表示した．

これを見ると，将棋 81 万だけで訓練したプレイヤ（Shogi816K）は，通常初期局面だけで訓練したプレイヤ（Base）に比べて，振り飛車（中飛車・四間飛車・三間飛車・向飛車）は比較的得意だが，居飛車の角換わり・相掛かり・横歩取りの 3 戦法を苦手としていることがわかる．

これに対して，将棋 81 万を事前学習で用いたプレイヤ（Shogi816K+Base, Shogi816K+Random2ply）では，振り飛車系が強いという性質を持ちつつ，さらに居飛車の将棋でも Base よりも全体的に強くなっている．

将棋 81 万のみで学習させるよりも，将棋 81 万を事前学習してから通常の将棋に近い局面を学習させるほうが，様々な戦型で棋力が向上するという結果になった．事前学習・適応学習の手法は，将棋の強化学習でも有効だったと考えられる．

5.3 将棋 81 万における先手・後手の駒配置の対称性が与える影響

本論文の「将棋 81 万」では，後手の駒は先手の駒の「回転対称」となるように配置することを提案した．

しかし，後手の駒を「回転対称」に配置するのが良いかは必ずしも自明ではない．そこで，将棋 81 万の対称性の違いで，以下の 3 種類のプレイヤを作成し，戦型別の棋力を測定した．ここでは将棋 81 万の対称性の違いがもたらす影響に焦点を当てるため，以下の 3 種類のプレイヤの作成にあたっては，事前学習後の適応学習は実施しなかった．

- **Shogi816K**: 回転対称の将棋 81 万を用いた自己対局のみで 100%学習．前述の Shogi816K と同じもの．（例：先手の玉が 1 九の枡に配置されたら，後手の玉

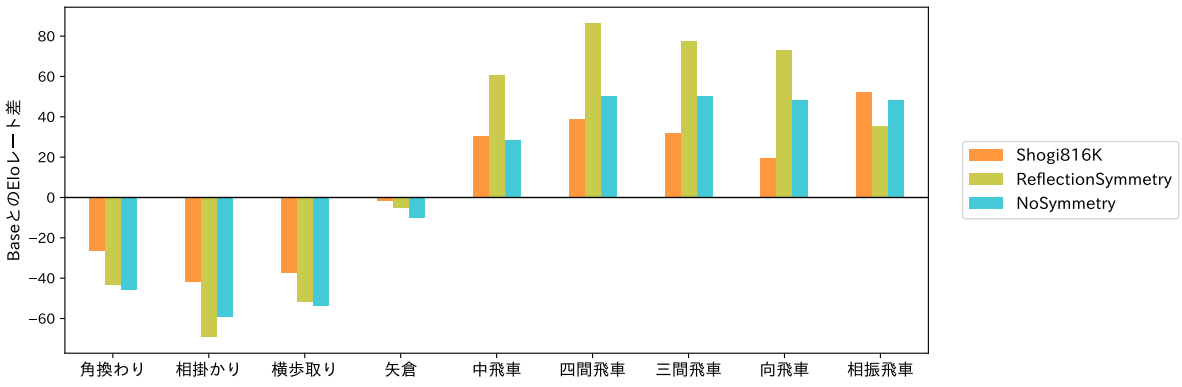

図 4 将棋 81 万における先手・後手の駒配置の対称性が与える影響．通常初期局面のみを学習したプレイヤ (Base) との Elo レート差を戦型別に測定．回転対称（Shogi816K），線対称（ReflectionSymmetry），対称性なし（NoSymmetry）を比較．

は９一の枡に配置）

- **ReflectionSymmetry**: 上記の将棋 81 万とは異なり，後手の駒を線対称に配置したもの．この初期局面集のみを用いた自己対局で 100%学習．（例：先手の玉が１九の枡に配置されたら，後手の玉は１一に配置）
- **NoSymmetry**: 将棋 81 万の先手の駒と後手の駒の対称性をなくし，後手の駒を先手の駒とは関係なくシャッフルして配置した初期局面集．この初期局面集のみを用いた自己対局で 100%学習．（例：先手の玉の位置に関係なく，後手の玉は１一から９一の枡のいずれかにランダムに配置）

なお，学習コストの観点で公平に比較するため，いずれの環境も自己対局数は自己対局数込みで合計 1000 万対局に統一した．

これらのプレイヤの戦型別の棋力を測定した結果が図 4 である．図 4 では，通常初期局面のみで訓練したプレイヤ（Base）を± 0 Elo として表示している．

まず，これらの 3 種類のプレイヤを通常初期局面のみで学習したプレイヤ（Base）と比較すると，居飛車に弱く振り飛車に強いという将棋 81 万の全体的な傾向は類似している．

他方で，対称性による得意・不得意の違いも見て取れる．居飛車の将棋では，回転対称の将棋 81 万で学習したプレイヤ（Shogi816K）が比較的強く，振り飛車の対抗形では線対称の将棋 81 万で学習したプレイヤ（ReflectionSymmetry）が強い傾向がある．線対称の将棋 81 万では，先手と後手の飛車が常に向かい合うように配置されるため，対抗形に適した特徴をニューラルネットワークが学習した可能性がある．

そして，対称性がないバージョン（NoSymmetry）は中間的な結果となった．

以上の結果からすると，対抗形の振り飛車を強くしたい場合は線対称の将棋 81 万を用いるのが良いが，居飛車の頻度が高い現代将棋の一般的な傾向に合わせる場合は，居飛車に比較的強い回転対称の将棋 81 万が全体的な棋力では有利になりそうである．

5.4 将棋 81 万の初期局面における先手有利・後手有利の偏りの程度

将棋 81 万のように駒の初期配置をランダムにシャッフルした場合，通常の将棋では滅多に出現しない駒の配置が混ざることになる．このことだけを考えると，先手・後手のどちらかに大きく形勢が傾いた局面が多い可能性もある．

この点を調べるため，これまでの対局実験とは別に，将棋 81 万の初期局面の評価値の分布を水匠で評価させた．ここでいう「評価値」とは，その局面で先手・後手どちらが勝ちやすいかをプログラムが推定した指標であり，プラスになると先手有利，マイナスだと後手有利，± 0 だと互角という意味である．一般的なプログラムでは，将棋で一番価値の低い駒とされる「歩」を先手が 1 枚分得した局面で+100 点となるように調整されている[*6]．

結論から言えば，将棋 81 万の初期局面は，81 万通り以上の多数の局面数を有しつつも，水匠の評価で先手・後手のどちらかに極端に有利になっている局面の割合は多くはないことがわかった．

図 5 は，将棋 81 万の初期局面 816,480 局面の中から一様乱数でサンプリングした 3,000 局面を，水匠 5 で各 10,000 ノード探索した場合の評価値の分布である．また，比較対象として，既存の将棋の棋譜から人間が設けた一定の基準で互角局面を選別した互角局面集[*7]から一様乱数でサンプリングした 3,000 局面の評価値の分布と，通常初期局面

[*6] 水匠を含む最近の将棋プログラムは USI (Universal Shogi Interface) プロトコル (`http://shogidokoro.starfree.jp/usi.html`) に準拠したものが多いため．

[*7] 2023 年の世界コンピュータ将棋選手権で優勝したプログラム dlshogi の作者である山岡忠夫氏が主にコンピュータ同士の対局用・棋力評価用として公開している中盤互角局面集．`https://tadaoyamaoka.hatenablog.com/entry/2022/12/31/114258`

から 2 手または 4 手一様ランダムに進めた時の局面を各 3,000 回復元抽出でサンプリングした評価値の分布を並べて掲載した．

また，表 1 は，各局面集の総局面数と，図 5 の評価値分布の平均・標準偏差をまとめたものである．

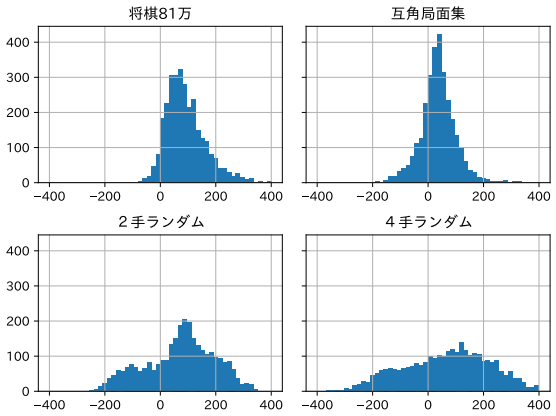

図 5 水匠 5 で測定した各初期局面の評価値の頻度分布（各 3000 サンプルの評価結果）．横軸が評価値，縦軸が頻度を表す．プラスの評価値は先手有利，マイナスの評価値は後手有利，0 は互角を示す．局面集ごとに総局面数は異なるが，サンプル数を揃えるため，各局面集について 3000 回ずつ復元抽出で一様ランダムサンプリングして評価．

	総局面数	評価値の平均	評価値の標準偏差
将棋 81 万	816,480	91.8	75.1
互角局面集*7	8,187	37.6	76.5
2 手ランダム	900	71.9	128.3
4 手ランダム	719,731	78.6	218.1

表 1 各局面集の比較．評価値の平均と標準偏差は図 5 の分布から求めたもの．

まず総局面数を確認すると，816,480 通りの局面を有する将棋 81 万が最も多い．今回比較対象とした互角局面集は 8,187 局面であるが，今回のものを含むこれまでの一般的な互角局面集は強いプレイヤ同士の既存の棋譜から局面を選別して作成されるため，棋譜を増やすか選別基準を緩めなければこれ以上局面数を増やすことは難しい．今回の互角局面数も数万局単位の棋譜から選別して作られている．

次に将棋 81 万の初期局面の評価値を見ると平均 91.8 点であり，今回比較した中ではやや先手有利の傾向があった．しかし，将棋はもともと先手やや有利のゲームであり，評価値の絶対値が 100 点程度まであれば一般的には互角の範囲内と言われることが多い．実際，今回の互角局面集*7 も評価値が± 170 点以内という条件で選別した局面を互角局面として扱っている．そのため，将棋 81 万の初期局面は平均的に十分互角の範囲と考えられる．

そして将棋 81 万に含まれる局面の評価値の標準偏差は 75.1 点であった。これは互角局面集と同程度であり．2 手ランダム（標準偏差 128.3 点）や 4 手ランダム（標準偏差 218.1 点）よりも評価値のばらつきが少ない．ランダムに数手指した局面を初期局面とした場合は初手で取る手や王手などが含まれうるが，将棋 81 万ではそうした手がないような駒配置の制約があるため，将棋 81 万では極端に有利不利が生じる局面が少ないと考えられる．

このように，将棋 81 万の利点としては，既存の棋譜を用いない比較的簡単な方法で多数の局面を持った初期局面集を作成できる点がある．また，将棋 81 万の局面のほうが数手ランダムに指す場合よりも極端な先手後手の有利不利が生じにくい利点があることがわかった．したがって，既存の棋譜を用いずに自己対局による強化学習でエージェントを訓練する際には将棋 81 万が適していると考えられる．

6. まとめと今後の展望

我々は，より多様性を持った将棋の初期局面集として将棋 81 万を考案した．将棋 81 万の多様な初期局面で事前学習を行なったエージェントは，そうでないエージェントに比べて，様々な戦法で+20 Elo から+50 Elo 程度棋力が向上した．特に，AlphaZero 型の強化学習では比較的学習が難しいとされる振り飛車戦法での効果が大きく，+50 Elo 程度の改善が見られた．多様性を持った環境で学習することが，様々な戦法での棋力向上に寄与したと考えられる．

提案手法の利点としては，特定の強化学習アルゴリズムに依存しないため，様々な強化学習アルゴリズムと併用可能である点が挙げられる．また，将棋 81 万の初期局面は，駒の初期配置をシャッフルするだけで準備できるため，比較的実装が容易という利点も備えている．

近年，将棋観戦や将棋研究に将棋 AI が活用されつつある [15]．提案手法は，より多様な戦法で AI の形勢判断を改善することに繋がり，人間の将棋観戦や将棋研究に役立てられると考えられる．

将来の展望として，本研究類似の手法を他のゲームに応用することが考えられる．本研究の将棋 81 万に対応するものとして，チェスでは既にチェス 960 が存在する．また，他のゲームでも初期状態に一定のランダム性を加えることは十分想定できる．将来的には，チェスやオセロなどの他のゲームでも，より多様な初期局面を用いて事前学習を行うことで，エージェントのパフォーマンスや頑健性が向上する可能性がある．

参考文献

[1] International Chess Federation: FIDE Handbook, (online), available from ⟨https://handbook.fide.com/chapter/E012018⟩.

[2] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M. and Bolton, A.: Mastering the game of go without human knowledge, *Nature*, Vol. 550, No. 7676, pp. 354–359

(2017).
[3] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K. and Hassabis, D.: A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play, *Science*, Vol. 362, No. 6419, pp. 1140–1144 (2018).
[4] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K. and Ostrovski, G.: Human-level control through deep reinforcement learning, *Nature*, Vol. 518, No. 7540, pp. 529–533 (2015).
[5] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D. and Wierstra, D.: Continuous control with deep reinforcement learning, *International Conference on Learning Representations* (2016).
[6] Cobbe, K., Hesse, C., Hilton, J. and Schulman, J.: Leveraging Procedural Generation to Benchmark Reinforcement Learning, *International Conference on Machine Learning*, PMLR, pp. 2048–2056 (2020).
[7] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W. and Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world, *International Conference on Intelligent Robots and Systems*, IEEE, pp. 23–30 (2017).
[8] 篠田正人：将棋における実現可能局面数について，ゲームプログラミングワークショップ 2008 論文集，pp. 116–119 (2008).
[9] Nakayashiki, T. and Kaneko, T.: Maximum Entropy Reinforcement Learning in Two-Player Perfect Information Games, *IEEE Symposium Series on Computational Intelligence*, Orlando, FL, USA, IEEE, pp. 01–08 (2021).
[10] Zahavy, T., Veeriah, V., Hou, S., Waugh, K., Lai, M., Leurent, E., Tomasev, N., Schut, L., Hassabis, D. and Singh, S.: Diversifying AI: Towards Creative Chess with AlphaZero (2023). arXiv:2308.09175 [cs].
[11] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, *Advances in Neural Information Processing Systems*, Vol. 25, Curran Associates, Inc. (2012).
[12] 岡谷貴之：深層学習（改訂第 2 版），講談社 (2022).
[13] Laskin, M., Lee, K., Stooke, A., Pinto, L., Abbeel, P. and Srinivas, A.: Reinforcement learning with augmented data, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 19884–19895 (2020).
[14] Danihelka, I., Guez, A., Schrittwieser, J. and Silver, D.: Policy improvement by planning with Gumbel, *International Conference on Learning Representations* (2021).
[15] 千田翔太：進化し続けるコンピュータ将棋：3. プロ棋士から見たコンピュータ将棋の活用，情報処理， Vol. 59, No. 2, pp. 157–160 (2018).
[16] Loshchilov, I. and Hutter, F.: Decoupled Weight Decay Regularization, *International Conference on Learning Representations* (2018).
[17] Wu, D. J.: Accelerating self-play learning in go, *arXiv preprint arXiv:1902.10565* (2019).

付　　録

A.1　強化学習の実験条件

本研究において AI エージェントを強化学習で訓練する際に用いた設定等は，表 A·1 のとおりである．

ニューラルネットワークの構造については，計算量削減のため，文献 [14] を参考に 16 ブロックの Bottlenecked ResNet を採用した．チャネル数は基本 128 チャネル，ボトルネック部は 64 チャネルとなっている．8 ブロック毎にプーリングブロック [14], [17] を含む．

なお，今回の実験では，千日手で引き分けになった棋譜を学習に含めると学習が不安定になったため，千日手となった対局は学習からは除くこととした．

自己対局の設定	1 手 16 シミュレーション
PUCT の定数	$c_{base} = 19652.0, c_{init} = 1.25$ [3]
最適化手法	AdamW [16]
学習率	0.001
L2 正則化係数	1e-4
Training window size	1,000,000 局
Games per checkpoint	20,000 局
並列探索	訓練時・評価時とも未利用
ネットワークの重み初期化	PyTorch 2.0 のデフォルト
Batch Normalizaton の設定	PyTorch 2.0 のデフォルト

表 A·1　強化学習時の設定

A.2　学習時の棋力の推移

学習途中の各エージェントの棋力の推移を計測したものが，図 A·1 である．図 A·1 を見ると，最初のうちは，通常初期局面だけで学習した場合（Base）の方が将棋 81 万のみの場合（Shogi816K）よりも早く学習が進んでいる．しかし，将棋 81 万で多様な局面を事前学習したエージェント（Shogi816K+Base, Shogi816K+Random2）は，通常初期局面への適応学習が始まると棋力を伸ばし，最終的に通常初期局面だけで学習したエージェント（Base）を追い抜いている．

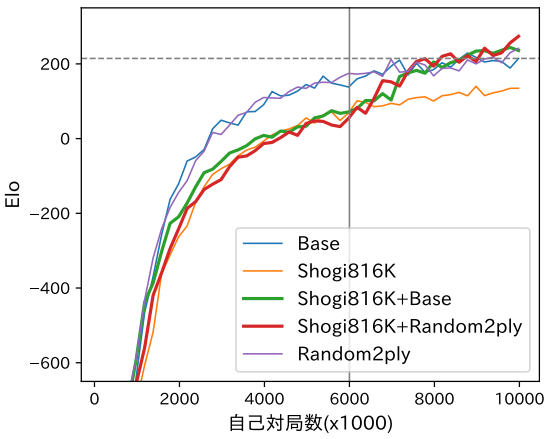

図 A·1　水匠の標準定跡で測定した学習中の棋力の推移．横軸が自己対局の対局数，縦軸が水匠（1000 ノード）との Elo レート差を示す．最終的な Base の棋力 (+214 Elo) に水平の点線を引いた．また，事前学習から適応学習に切り替わるタイミング（60%時点）には垂直線を引いた．