強化学習を用いて人間の直感を再現した詰将棋解答 AI の作成

千葉　景太[1,a)]　Reijer Grimbergen[2,b)]

概要：2013 年，公式戦で将棋のプロが初めて将棋 AI に敗れた．囲碁においても人間より強いものが開発された．これらの理由から完全情報ゲームにおける強い AI の研究は概ね決着がついたと思われる．しかし，人間らしい AI の研究はまだ終わっていない．コンピュータが人間と同じ局面数で，人間と同程度の指し手を見つけるのは困難である．そこで，本研究では AI による局面の探索を行わずに指し手の精度を上げることで，人間の直感を再現した AI の作成を目指した．AI はポリシーネットワークに従って指し手を決定し，指し手の正解率を上げるために学習を行った．詰将棋が学習に用いる局面はランダムプレイヤー同士の対局で現れた，詰む局面を用いた．その結果，ランダムプレイヤーを用いて生成した局面であれば 60%程度の精度で正解手を指すようになった．しかし，本来の詰将棋の局面に対しては，AI の性能を向上させることはできず，王手のみを抽出したランダムな指し手よりも低い精度で正解手を指すようになった．

Research into Using Reinforcement Learning to Reproduce Human Intuition in a Shogi Mating Problem Solver

Keita Chiba[1,a)]　Reijer Grimbergen[2,b)]

Abstract: In 2013, a shogi professional was beaten for the first time by a shogi AI in an official match. Also in Go, programs that are stronger than the best human players have been developed. Because of this, it can be said that for complete information games, the research into the building of strong AI programs has almost come to an end. However, research into human-like AI has not finished. It is a difficult problem to build an AI that can select similar moves to human expert players while only considering a limited amount of positions. Therefore, our research aims at building an AI that makes good move decisions without doing any search, thereby simulating human intuition. The proposed AI makes move decisions using a trained Policy Network, using reinforcement learning to improve the times a correct move was selected by the network. For learning, mating situations from games between AI playing random moves were used. As a result of the training, in about 60% of these random mating positions, the correct move was selected. However, this did not carry over to proper mating problems, where the proposed method was outperformed by selecting a checking move randomly.

1 東京工科大学大学院 バイオ・情報メディア研究科 コンピュータサイエンス専攻
2 東京工科大学 コンピュータサイエンス学部
a Tokyo University of Technology, Graduate School of
Bionics, Computer and Media Science, Entrepreneurship Program
b Tokyo University of Technology
School of Computer Science

1. はじめに

2013年, 公式戦で初めて将棋のプロが将棋 AI に敗れた. また，2017年に人間の棋譜を一切使わずに難易度が高いとされていた囲碁において，人間以上の強さの AI が開発された[1].

以上より，完全情報ゲームにおける強い AI の研究はほぼ決着がついたと思われる．しかし，人間らしい手を指す AI の研究はまだ終わっていない．将棋 AI は一秒あたり数万以上の局面を探索可能である．一方，人間はトッププロでさえ一分間に20〜30局面しか読むことができない[2]．現在トップレベルの AI でも，この探索局面数でトッププロと同程度の強さの指し手を見つけるのは困難である．AlphaGo[3]のポリシーネットワークは教師あり学習を行い，専門家の指し手と57%一致させた．そこで，本研究では AI による局面の探索を行わずに指し手の精度を上げることで，人間の直感を再現した AI の作成を目的とする．

囲碁や指し将棋は，指し手の正解と不正解の判断が難しく，AI の指し手を正しく評価するのは困難である．したがって，本研究では詰将棋を用いて AI の性能を評価する．

人間が詰将棋を解くとき，勘で次の一手をある程度予測することができる．例えば，図1のような詰将棋を解くときに，人間は最後まで読み切らなくても赤い矢印のような手が良さそうであると予測できる．また，合法手を列挙し，2017年の世界コンピュータ将棋選手権で優勝した AI である elmo の評価関数を使って評価すると，青の矢印の手の評価値が最も低く, 緑の矢印の手の評価値が最も高くなる．

本研究では実験の事前データとして elmo の評価関数を用いる．詰将棋の初手を elmo を使って評価した結果と初手をランダムで選択したときの正答率を表 1 に示す．3 手詰は 3488 問，5 手詰は 3533 問のデータを用いた．このときの手数は攻方と玉方が最善手を選択したときの詰むまでの手数であり，指し手は王手の生成は王手のみとした．

3手詰と5手詰の正答率に大きな差はなかった．また，elmo の評価関数とランダムプレイヤーの間にも大きな差はなく，elmo の評価関数を使っても正答率は低い．

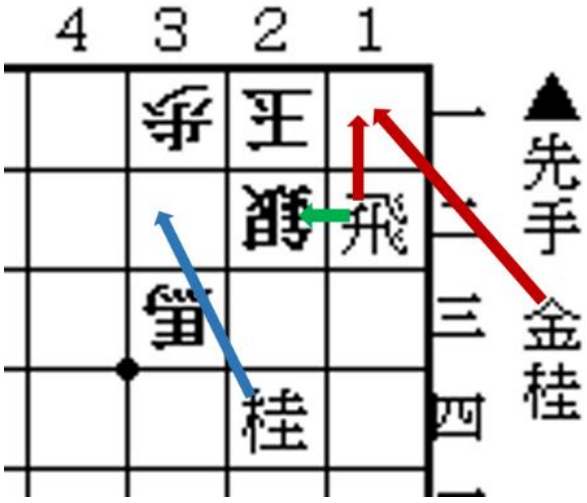

図 1 直感的な手とそうでない手の例

表 1 探索なしの初手の正答率

評価関数	3手詰	5手詰
ランダム	13.6%	14.2%
elmo	14.0%	18.3%

2. 提案手法

本研究では，探索を行わずに詰将棋を解答する AI を作成することで, 人間が直感で解答することの再現を目指す. AI の実装は文献[4]で解説されている将棋 AI を参考にした.

2.1 指し手の生成

AI の指し手は，AlphaGo でも用いられていたポリシーネットワークによって決定する.

ポリシーネットワークとは，局面を入力として指し手を予測するニューラルネットワークであり，出力は合法手の確率分布とする．このとき，ポリシーネットワークが正しく学習できていれば，出力される指し手のほとんどは王手になると考えられるため，出力は王手以外も含む.

2.2 局面の生成

学習とテストに使う詰将棋の局面は，ランダムプレイヤーの対局で現れた詰将棋の局面を利用する.

このとき，完全作の詰将棋だけで本研究の学習に必要な局面数用意するのは困難であるため不完全作のものも採用する．それにともない，詰将棋のルールを以下のように簡略化する.

- 攻方は最短で詰ます手を選ばなくても良い.
- 玉方は最長で逃げられる手を選ばなくても良い. しかし，最善手ならば逃げられる局面で詰まされる手を指してはいけない.
- 詰将棋の正解手順は一つでなくても良い．また，持ち駒が余っても良い.

2.3 ネットワークの構成

ネットワークの構成は Residual Network[5]を使用する. Residual Network は深いネットワークの学習に適してニューラルネットワークであり，本研究で用いるものは図 2 のとおりである．図 2 の左がネットワーク全体の構成，右側が各 Block の構成である．一層目に畳み込み層とし，二層目から Residual Block を 5 つ繰り返す．最後に畳み込み層の構成とし，Residual Block は二層の畳み込み層で構成する.

フィルターのサイズは 3×3 で，中間層のフィルターの枚数は 192，ストライドとパディングが 1 でプーリング層なし，活性化関数は ReLU とする.

入力は現局面とし，入力チャンネルのラベルを表 2 に示す．盤上の駒種と持ち駒の最大数と先手後手の合計で 104 チャンネルとする.

出力のラベルについて表 3 に示す．ラベル数は移動方向と駒の移動先の座標の組み合わせであり，出力ラベル数は 27×81=2187 となる.

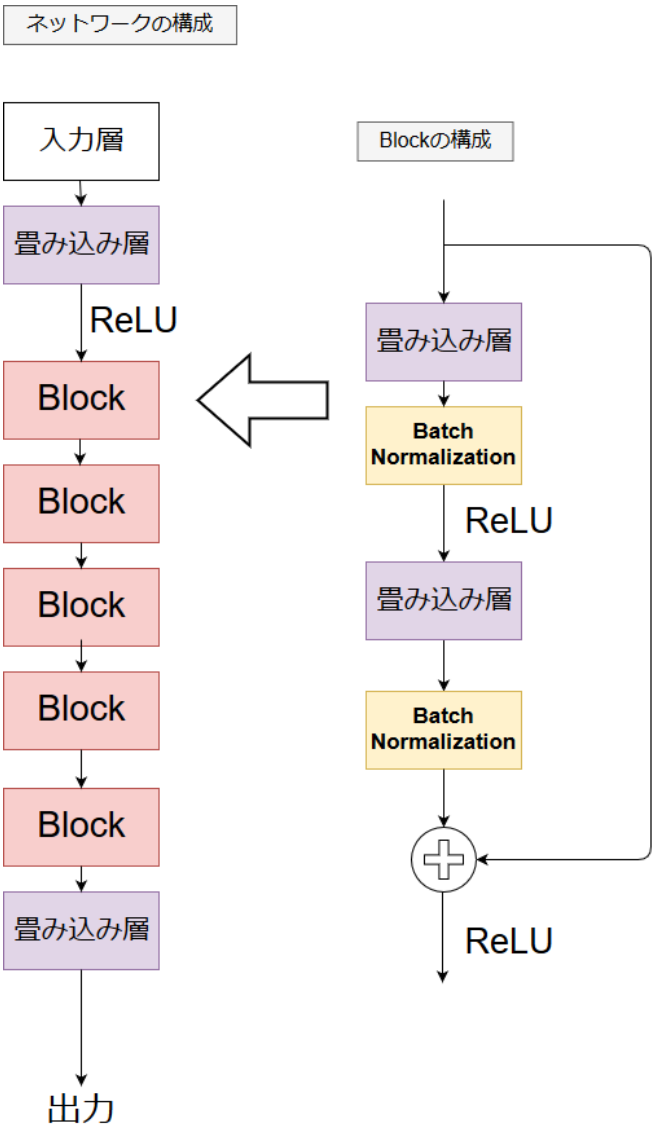

図 2　Residual Network の構成

表 2　入力チャンネル数

	チャンネル数
盤上の駒	14
持ち駒	38
合計	52
先手，後手合計	104

表 3　駒の移動に割り当てるラベル

種類	チャンネル数
移動する手	10
持ち駒から打つ手	7
成る手	10
合計	27

2.4 学習と評価方法

詰将棋の局面から，一手ポリシーネットワークに従って指し手を決める．その手で相手玉を詰ますことができるのであればその手を正解手としてポリシーネットワークを学習させる．また，正解手でない場合は正解手の中からランダムで一手選び，その手を正解手としてポリシーネットワークを学習させる．最適化アルゴリズムは AdaGrad[6]を用いた．　AI の精度の評価は，詰将棋の局面でポリシーネットワークが正解手を指した割合とし，指し手の選択には貪欲法を用いる．

3. 実験結果

本研究で作成した AI の詰将棋の正答率を図 3 に示す．事前データで用いた三手詰の詰将棋 3888 問の正答率がオレンジの線で，ランダムプレイヤーが生成した詰将棋 10000 問の正答率が青い線である．また，これらの詰将棋は学習には用いていない．60000 局面学習したときのポリシーネットワークの精度は約 60%となり，AlphaGo のポリシーネットワークと専門家の一致率とほぼ同じとなった．しかし，本来の詰将棋の正答率はほとんど上がらなかった．AI が王手をした確率を図 4 に示す．ランダムプレイヤーが生成した局面と本来の詰将棋の局面で，ともに 8 割程度となっている．このことから AI は，王手をしたほうが良いということは学習できているようだが，絶対に王手をしなければいけないということまでは学習できていない．また，王手率も 1 学習局面数が 10000 のときから，ほとんど上がっていないため，学習は不十分である．

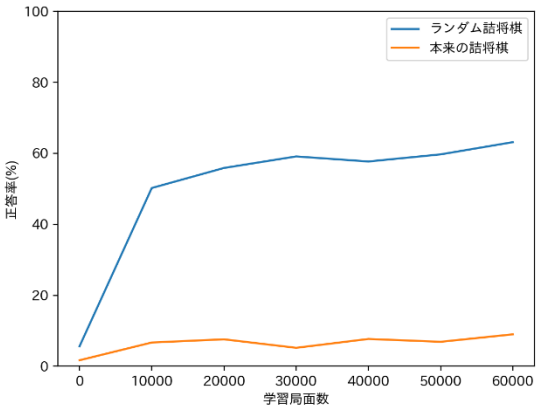

図 3　学習局面数と AI の正答率

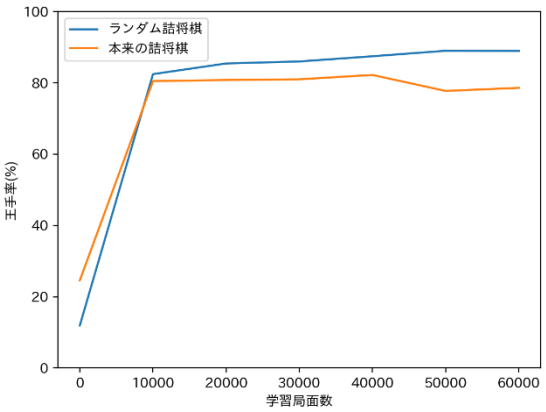

図 4　学習局面数と AI が王手した割合

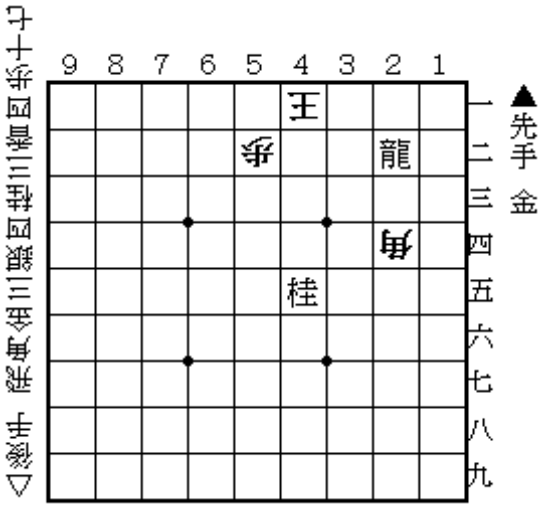

図 6　本来の詰将棋の例

4. まとめと今後の課題

本研究ではポリシーネットワークを強化学習して詰将棋を解答する AI を作成した．今回は学習する局面の生成にランダムプレイヤーを用いたため，図 5 のような局面が出来上がってしまった．しかし，本来の詰将棋は図 6 のようなものである．したがって，本来の詰将棋を解く感覚と，本研究で生成した局面を解く感覚がかけ離れてしまったと考えられる．そのため，ランダムプレイヤーが生成した局面の精度はある程度まで上がったにも関わらず，本来の詰将棋を解く精度は上がらなかった．しかし，今回の実験では，まだ正答率が飽和していないため，もうしばらく学習を続けたい．

今後の課題として，学習局面に使用する詰将棋を質の良いものや完全作の使用，ポリシーネットワークとバリューネットワークの転移学習，より良いネットワークの構成を用いることなどが考えられる．

図 5　ランダムプレイヤーが生成した局面の例

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 & Demis Hassabis, “Mastering the game of Go without human knowledge”, Nature 550, pp354-359(2017)

[2] 伊藤毅志，松原仁「先を読む頭脳」新潮社(2006)

[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, Demis Hassabis, “Mastering the game of Go with Deep Neural Networks and Tree Search”, Nature 529, pp484-489(2016)

[4] 山岡忠夫「将棋 AI で学ぶディープラーニング」株式会社マイナビ出版(2018)

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Deep Residual Learning for Image Recognition”, IEEE Conference on Computer Vision and Pattern Recognition, pp770-778 (2016)

[6] John Duchi, Elad Hazan, Yoram Singer, “Adaptive Subgradient Methods for Online Learning and Stochastic Optimization”, Journal of Machine Learning Research 12, pp2121-2159(2011)