F-003

コンピュータ囲碁の強化学習における着手限定ルールの追加による学習効率の評価
Evaluation of Learning Efficiency by Adding Move Restriction Rules in Reinforcement Learning for Computer Go

谷田 聖司† 小田 凌平† 藤田 玄†
Masashi Tanida Ryohei Oda Gen Fujita

1. はじめに

AlphaZero[1]のように完全情報ゲームにおいてルールを記述し、それを基に白紙の状態から強化学習を繰り返す学習方法が有効であることが知られている。ただし、コンピュータ囲碁においてこの手法を適用すると、特に学習初期において、ルールには従っているものの眼を潰すなどの不利な手を繰り返し、結果的にお互いに石を取り合うという事例が発生する。このような学習初期に見られる品質の低い学習データは学習効率に悪影響を及ぼしていると考えられる。そこで本研究では、本来のルールにはない、眼には着手しないというルールを追加することで着手を限定し、強化学習を行う手法を提案し、学習効率の評価を行った。その結果、本手法を適用した場合、学習効率の向上がみられ、既存学習結果に勝利するなど、棋力の向上が確認できた。

2. 囲碁における眼

眼とは、囲碁用語の一種であり、一色の石で囲まれた座標の事を指す。囲碁のルールにより、通常は相手の石で形成された眼に着手することは自殺手という手になり、禁止されている。よって、盤面上に自信の石で形成した眼があると有利な状態となる。しかし、相手の石が形成する眼の周りを自身の石で囲むと、眼に着手する事ができるようになり、眼を形成する石を取ることができる。また、眼を形成する石が連なり、二つの眼を形成する二眼と言う状態であれば、相手の二眼の周りを自身の石で囲っても眼に着手することは出来ない。眼の例を図 1 に示し、二眼の例を図 2 に示す。丸で示した座標が眼である。

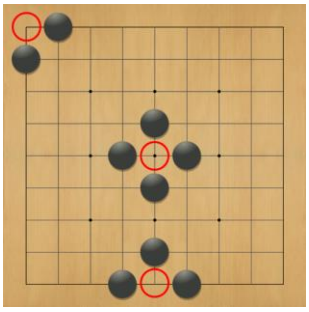

図 1　眼

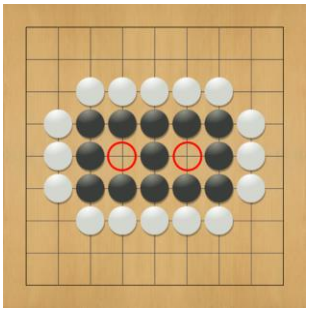

図 2　二眼

3. コウ

コウとは、着手することにより相手の石を取ることができた石を次の手で取り返す事を禁止するルールである。図 3 に示すように、白が石を 1 に着手し黒を取り、次の手で黒が 2 に着手し、白の石を取り返す事は出来ない。

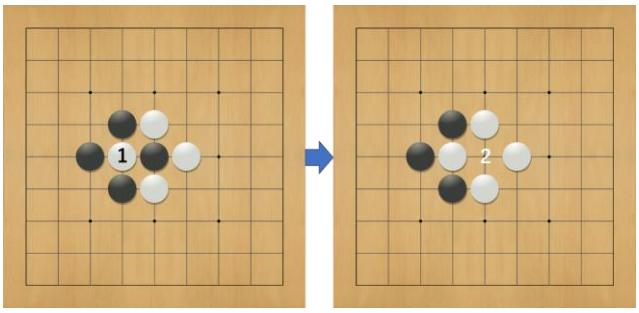

図 3　コウ

4. AlphaZero

AlphaZero とは DeepMind 社によって開発されたコンピュータプログラムである。囲碁プログラムの AlphaGo Zero を汎化させたものであり、将棋、チェス、囲碁において世界チャンピオンプログラムとなる。AlphaZero は人間が作成した棋譜データを用いず、自己対局によって学習する。自己対局はほぼランダムな着手から始まり、初期段階において棋力向上に時間がかかる。

5. 強化学習時の問題点

囲碁は自身と相手がお互いに連続してパスを行ったときに終局する。そこで強化学習の自己対局においてランダムな着手が続いた場合、終局に辿り着かないことがあり、学習効率が悪く、学習データに悪影響を及ぼすことがある。例えば、図 4 のように黒は眼を 2 つ形成しており、黒石は絶対に取られないという状態にあるが、ランダムな着手により黒は自身の眼に着手を行ってしまう。そして、白が空いている座標に着手を行うと黒は石を全て失う。

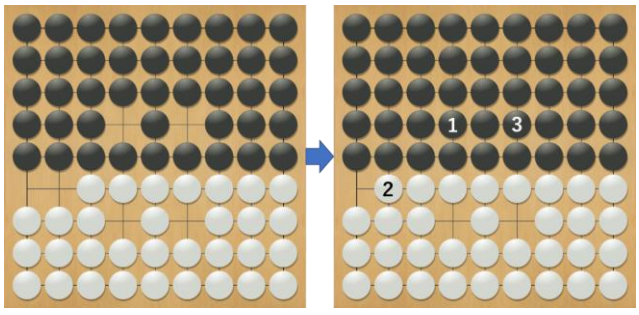

図 4　ランダム着手の不利となる着手

6. 強化学習時の着手限定

強化学習の自己対局において、ランダムな着手により終局に辿り着かず、質の悪い学習データが生成される事を防ぐために、眼には着手しないというルールを追加し、着手の限定を行う手法を提案する。本稿で用いる囲碁プログラ

† 大阪電気通信大学　Osaka Electro-Communication University

ムは、AlphaZero を参考に開発されたフリーソフト[2]である。この囲碁プログラムには、着手を行うときに、着手点が着手禁止点かどうかを判定するシミュレータ部分があり、そこに眼に着手しないルールを組み込んだ。囲碁ルールと追加ルールにより着手できない座標を着手禁止点とし、着手禁止点の判定を行う手順をフローチャートにより図 5 に示す。座標を盤面 $p_{x,y}$ とする。

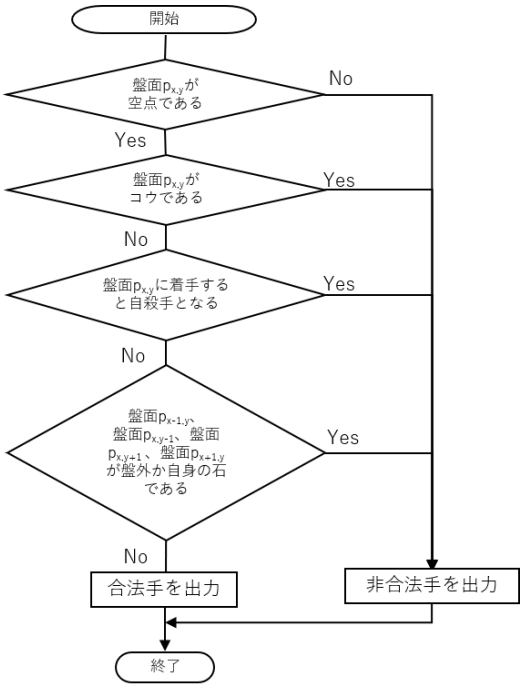

図 5　着手禁止点の判定

7. 性能評価

強化学習において、通常の囲碁ルールにより自己対局を行い、作成した学習データと、眼には着手しないというルールを追加して自己対局を行い、作成した学習データを用いて対戦をさせ、どちらのルールで作成した学習データの棋力が高いかを確かめた。通常の囲碁ルールによる学習データと追加ルールによる学習データの対戦は通常ルールにより行った。学習データは 4 時間、8 時間、12 時間学習させたものを用い、対戦数は 200 戦である。学習データ作成時の環境を表 1 に示す。対戦結果を表 2 に示す。結果は、着手を限定して学習した学習データが 9 割以上勝利した。通常のルールで学習させた学習データよりも、早く棋力が上昇したと分かった。

表 1　学習データ作成時の環境

OS	Windows10 Home
CPU	Intel® Core™ i7-9700K
GPU	NVIDIA GeForce RTX 2070 SUPER
RAM	32.0 GB
学習ツール	Tensorflow 1.9.0

表 2　対戦結果

学習時間	通常ルール	着手限定ルール
4 時間	20 勝	180 勝
8 時間	18 勝	182 勝
12 時間	16 勝	184 勝

8. おわりに

コンピュータ囲碁の強化学習の自己対局において、眼には着手しないというルールを追加して着手を限定することにより、学習効率を向上をさせることができた。

参考文献

[1] David Silver, David Silver, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepe, Timothy Lillicrap, Karen Simonyan, Demis Hassabis, “A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play”, Science, Vol. 362, Issue 6419, pp. 1140-1144 (2018).

[2] Surag Nair, “A clean implementation based on AlphaZero for any game in any framework + tutorial + Othello/Gobang/TicTacToe/Connect”, https://github.com/suragnair/alpha-zero-general (2021).