コンピュータ囲碁の強化学習における着手限定ルールに対する条件付けの検討

谷田 聖司　小田 凌平　藤田 玄

概要：AlphaZero 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。この問題に対し、著者らは強化学習における自己対局時に眼には着手しないというルールを追加する事により、初期段階において学習効率を向上させる手法を提案した。しかし、囲碁の対局では眼への着手を行った方が良い場合もあるという課題がある。そこで本稿では、眼への着手をした方が有利となる場合、眼への着手を可能とする手法を提案する。

A Study on Conditioning for Move Restriction Rules in Reinforcement Learning for Computer Go

Tanida Masashi　Oda Ryohei　Fujita Gen

Abstract: In perfect information games, it is known that a learning method in which reinforcement learning is repeated from a state of no learning using only the game rules is effective, and AlphaZero is a representative example. However, when this method is applied to computer Go, especially in the early stages of learning, there are cases where the players repeatedly make disadvantageous moves, such as moves that reduce their own liberties, and as a result, they end up taking stones from each other. Such low quality training data in the early stages of learning is thought to have a negative impact on learning efficiency. To solve this problem, the authors proposed a method to improve the learning efficiency in the early stage of reinforcement learning by adding a rule that no move should be made with less than two liberties in a self-playing game. However, there is a problem that it is sometimes better to make moves that reduce the liberties in a game of Go. Therefore, in this paper, we propose a method to enable moves to reduce the liberties with certain conditionalization.

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。

この問題に対し、著者らは強化学習における自己対局時に眼には着手しないというルールを追加する事により、初期段階において学習効率を向上させる手法 [2] を提案した。

しかし、囲碁の対局では眼への着手を行った方が良い場合もあるという課題がある。そこで本稿では、眼への着手をした方が有利となる場合、眼への着手を可能とする着手限定ルールに対する条件付けを提案する。

[1] Osaka Electro-Communication University

2. コンピュータ囲碁の用語

2.1 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、盤面上に自身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。

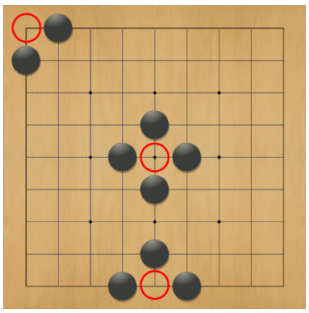

図 1 眼

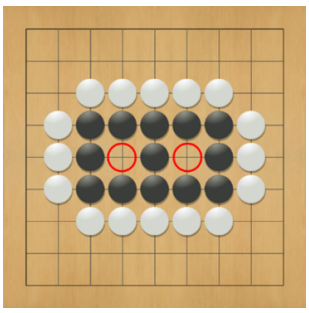

図 2 二眼

2.2 連

隣り合った同色の石の連なりのことを連と言う。連の情報がある事により、合法手判定の処理が早くなる。連の例を図 3 に示す。

2.3 呼吸点

石または連に隣接する空点のことを呼吸点と言う。呼吸点の数を数えることにより、相手にいくつ石を置かれると石が取られるかや、逃げ道の数かいくつあるかが分かる。1 つの赤丸が 1 つの呼吸点である。呼吸点の例を図 4 に示す。

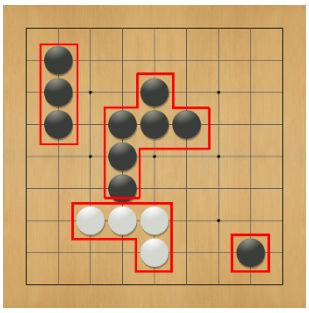

図 3 連

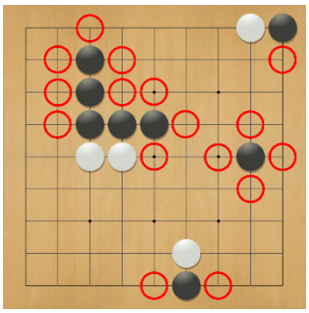

図 4 呼吸点

3. 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。ルールの記述部分と学習が分かれているため、様々な完全情報ゲームに適応できる。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。自己対局の流れを図 5 に示す。各手の探索を行い、強い手を選択する。本稿で用いる囲碁プログラムは、AlphaZero を参考に開発されたフリーソフト [2] である。このプログラムには囲碁ルールが実装されていないので、囲碁ルールの追加を行った。

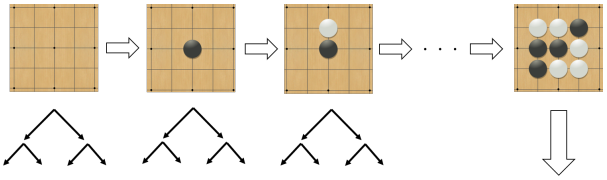

図 5 自己対局

4. 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。そして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 6 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。

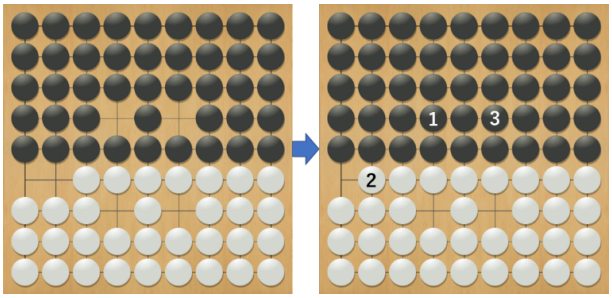

図 6　ランダム着手の不利となる着手

5. 強化学習時の着手限定 (従来手法)

5.1 概要

従来の手法 [3] では、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する。

5.2 実装

囲碁プログラムの AlphaZero 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ルールの記述部分があり、そこに、眼に着手しないルールを組み込む。眼の判定は、着手点の上下左右の座標全てに自身の石が存在するかしないかを参照し、存在した場合、眼であると判定する。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 7 に示す。囲碁の盤面の座標を盤面 $P_{x,y}$ とする。

5.3 従来手法の問題点

従来手法では、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たが、眼に着手しなければ不利となってしまう場合がある。例として、図 8 の目を作る左上の黒の連のように、呼吸点が 1 となっていた場合、黒は眼に着手しなければ相手は眼に着手でき、眼に着手されてしまうと眼を作る左上の黒の連の石は取られ、不利となる。このように、必ずしも眼への着手を禁止すれば良いというわけではない事が分かる。

6. 提案手法

従来手法の問題点に対処するため、従来手法に新たにルール追加を行う。従来手法の着手点が眼であるかどうかの判定を行った後に、その眼を作る石の連の呼吸点の数が 1 であるかどうかを判定する。1 であった場合、眼への着手を可能とする。これにより、眼への着手をしないと不利となってしまう場合に、着手が可能となる。眼を作る石の連の呼吸点の数による判定を追加した着手禁止点の判定の手順をフローチャートにより図 9 に示す。

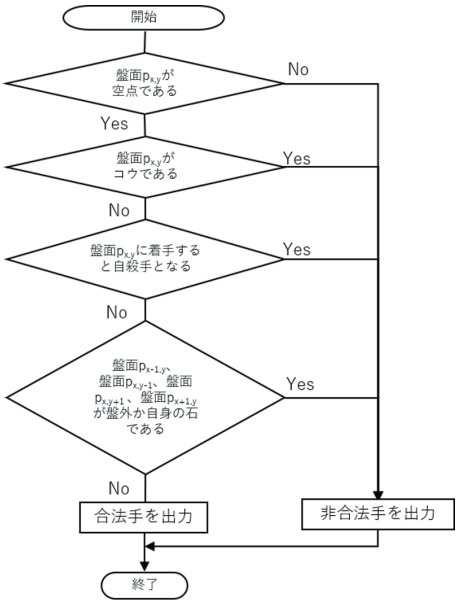

図 7　着手禁止点の判定

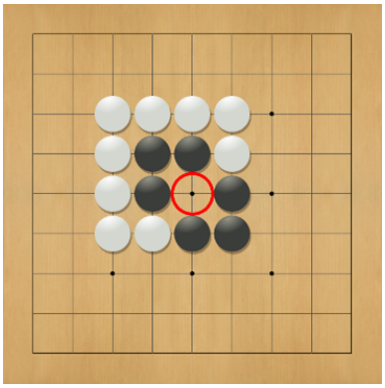

図 8　従来手法の問題点

7. 評価

提案手法の評価を従来手法の評価結果と比較し行った。具体的には、従来手法の着手制限ルールで学習した学習データと通常ルールで学習した学習データの対戦による勝率に対し、提案手法の着手制限ルールで学習した学習データと通常ルールで学習した学習データの対戦による勝率を比較した。学習データはそれぞれ 240 分行ったものを使用し、対戦は 200 戦行った。結果を表 1 に示す。提案手法は、従来手法において制限されていた有利となる着手を可能にし、従来手法よりも高い棋力が期待できる手法であったが、従来手法に比べて勝率が大きく下がった。勝率が下がった原因に対し、今後検討を行う。

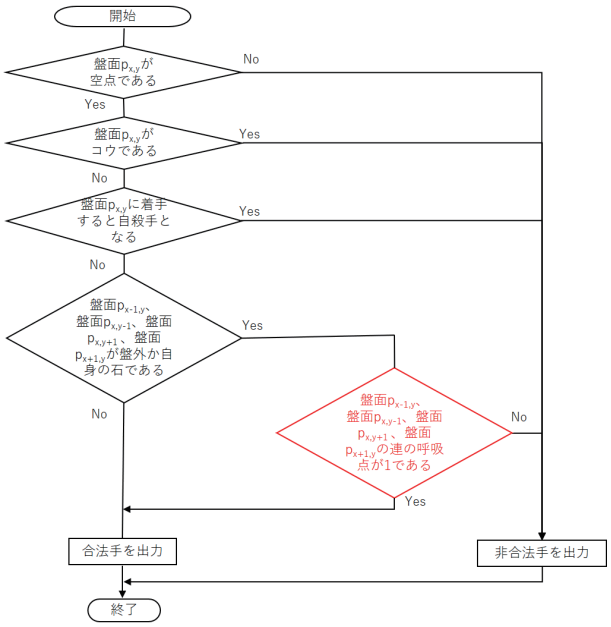

図 9 眼を作る石の連の呼吸点の数による判定を追加した着手禁止点の判定

表 1 通常ルールで学習した学習データとの対戦結果型

学習データ	勝	負	勝率 (%)
従来手法	188	12	94
提案手法	45	155	23

8. まとめ

囲碁プログラムの強化学習の初期段階において、自身の眼への着手により、終局せず、学習効率に悪影響を与えている問題があった。自身の眼への着手を制限することにより学習効率を向上させるという従来手法があったが、眼への着手をしないと不利となる場合があった。そこで、眼を作る連の呼吸点が 1 のとき眼への着手を可能とするルールを加える改良を提案した。

参考文献

[1] Silver, D. et al.: *A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play*, Science , 362(6419):1140-1144, 2018..

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

[3] 谷田 聖司, 小田 凌平, 藤田 玄： 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. 第 20 回情報科学技術フォーラム (2021).

