2019.6 のブログ：「最強囲碁ＡＩ　アルファ碁解体新書」を読んで
（→　http://www.1968start.com/M/blog/index.html#1907　）の別紙

「最強囲碁ＡＩ　アルファ碁解体新書」を読んで

中所　武司

■本考察のきっかけ

プロ棋士よりも強くなったアルファ碁（コンピュータ囲碁プログラム）について、
画像認識に効果的なディープラーニング（深層学習）の技術を適用との説明に、
画像認識と囲碁の盤面認識とは似て非なるものなので、疑問に思っていた。

画像認識では、類似の画像を同じカテゴリと判断できるが、
囲碁では、一路ずれると状況は大きく異なる場合が多いので、
類似の盤面で同じ出力でよい、とはいえない。

最近、以下の著書を読み、疑問が解けた。
・大槻知史：「最強囲碁ＡＩ　アルファ碁解体新書」、翔泳社、2017.

以下のような、囲碁に固有の処理に納得：

■注目した記述に関するメモ

【Chapter1　アルファ碁の登場】

【Chapter2　ディープラーニング　～囲碁 AI は瞬時にひらめく～】

●2.3　アルファ碁における畳み込みニューラルネットワーク

・2.3.2　「次の一手」タスクと画像認識の類似性

＊「次の一手」タスクと画像認識は似ている：
　表 2.1「手書き数字認識と囲碁ＡＩのＣＮＮの類似性
　→手書き数字を入力して、０～９のいずれかの数字を出力
　→囲碁の盤面情報を入力して、19*19 路の 361 種類の位置のいずれかを出力
＊大きな違いは：
　→手書き文字では、ピクセルの明るさを 256 段階で入力
　→ＣＮＮ（13 層）では、19*19 の路位置情報を入力
　　各位置の情報は 48 種類（値は０，１）
　　（例）黒石の位置、白石の位置、石の取れる位置、シチョウ、・・・

★コメント（再掲）：
　画像認識では、類似の画像を同じカテゴリと判断できるが、
　囲碁では、一路ずれると状況は大きく異なる場合が多いので、
　類似の盤面で同じ出力でよい、とはいえない。

＊「次の一手」タスクを実行するＣＮＮは、ＳＬポリシーネットワークとよばれ、
（ＳＬ：Supervised Learning 教師付き学習）
局面を入力して、各位置に打つ確率の予測値を出力する。→図 2.18：出力例

・2.3.4　ＳＬポリシーネットワークの入力４８チャネルの特徴

＊表 2.2　ＳＬポリシーネットワークの４８チャンネル分の入力
黒石の位置／白石の位置／空白の位置
ｋ（1～8）手前に打たれた位置
石がある場合の当該連の呼吸点（あと何手で取られるか）の数（1～8）
そこに打った後、石を取れるか（取る数ｋ：1～8）
そこに打った後、当該連を取られる場合に、何個石を取られるか（1～8）
その石に打った後の、当該連の呼吸点の数（1～8）
そこに打った後、隣接する相手の連をシチョウでとれるか
そこに打たれた後、隣接する味方の連をシチョウで取られるか
合法手か
すべて１で埋める／すべて０で埋める

・2.3.6　ＳＬポリシーネットワークの計算量

＊図 2.21　ＳＬポリシーネットワークの畳み込み計算とその計算量
パラメータの数：約４００万個
3**2*（フィルタの種類：192）**2*（層の数：12）

・2.3.7　ＳＬポリシーネットワークの学習用データの獲得

＊４００万個もあるフィルタ重みパラメータを学習するには
大量の高品質な学習データ（入力局面と正解ラベルの組）が必要

★コメント：
私の 50 年近く前の修士論文では、パラメータセットの値を変えて、いい結果を掲載した。
似ているかも（^^;;
http://www.1968start.com/M/bio/olduniv/7012sigA/7012SigAutomaton.pdf

・2.3.10　局面の有利不利を予測するＣＮＮ（バリューネットワーク）

＊ＳＬポリシーネットワークはある盤面に対して、各位置に石が打たれる確率を出力したが、
バリューネットワーク（ＣＮＮ）は入力局面の勝率予測値を出力
（アルファ碁の画期的な成果の一つ）
＊図 2.26　アルファ碁において勝率予測値を計算するバリューネットワーク

【Chapter3　強化学習　～囲碁 AI は経験に学ぶ～】

●3.4　迷路を解くための強化学習

・3.4.2　価値ベースの手法：Q学習により迷路を解く

＊Q学習は、ある行動を採るたびに、次に行くマスの価値と今いるマスの価値の
　差分を計算します。そしてその差分だけ、今いるマスの価値を増やす手法
＊例：図3.8　迷路のQ学習の結果

・3.4.3　方策ベースの手法：方策勾配法により迷路を解く

＊Q学習との違いは、価値関数の代わりに、各行動を採る確率を使用。
　ゴールに到達したときの行動の確率を少し高め、それ以外の行動の確率を少し下げる。
　ゴールした経路に含まれる行動は「良い行動であることが多い」という経験則。
＊例：図3.10　迷路の勾配方策法による結果

★コメント：
　教師あり学習でのニューラルネットワークのノード間の重み調整と類似か！？

●3.6　アルファ碁における強化学習

・3.6.1　アルファ碁の強化学習

＊図3.12　ＲＬポリシーネットワークを獲得するための強化学習の概要
　・強化学習の目的：ＳＬポリシーネットワークを強化学習することで、
　　より勝ちやすいＲＬポリシーネットワークをつくる
　　（ＲＬ：Reinforcement Learning）
　・ＳＬポリシーネットワークを初期値とし、「ゲームの勝利」を報酬として、
　　方策勾配法により強化学習
　　＃勝ったときは、勝つにいたる手をできるだけ選ぶようにパラメータ更新
　　＃負けたときは、負けるに至る手をできるだけ避けるようにパラメータ更新
　・自己対戦により、ゲームの結果を得る処理が膨大な時間を要するため、
　　学習には％ＧＰＵでも約１日かかる。

・3.6.2　方策勾配法に基づく強化学習

＊ＲＬポリシーネットワークを獲得するための手法

＃ステップ１：ＲＬポリシーネットワークのパラメータを初期化

＃ステップ２：相手モデルを更新
　過去の集合からランダムに選択し、自己対戦128回に１回の更新

＃ステップ３：相手モデルと味方モデルで終局まで手を進める（所要時間が圧倒的に大）
　この相手モデルにたいして、最新の味方モデルとの自己対戦を実施。128回を１セット。

＃ステップ４：方策勾配法により、ポリシーネットワークのパラメータを更新
　128回の自己対戦終了後、勝ち負け情報と、ＣＮＮの誤差逆伝搬法による勾配情報を元に、
　迷路の場合と同様に、ポリシーネットワークのパラメータを更新

＃ステップ５：ポリシーネットワークを相手モデルの集合に追加
　相手モデルのバリエーションをふやすため、128回の自己対戦を500セット繰り返すたびに
　そのときのポリシーネットワークを相手モデルの集合に追加

【Chapter4　探索 ～囲碁 AI はいかにして先読みするか～】

●4.4　囲碁におけるモンテカルロ木探索

・4.4.4　モンテカルロ木探索の結果と最終的な手の探索

＊モンテカルロ木探索では、探索結果をもとに、勝率に基づいて次の一手を選択。
　図 4.16　モンテカルロ木探索（最終的な手の選択処理）

【Chapter5　アルファ碁の完成】

●5.1　アルファ碁の設計図

・5.1.2　全体を制御するＡＩ

＊アルファ碁は、モンテカルロ木探索以外にも、ポリシーネットワークと
　バリューネットワークという素晴らしい評価指標を手にしました。
　これらをバランスよく使って全体を制御することが重要
　図 5.2　囲碁ＡＩの進化と「全体を制御するＡＩ」の役割の変遷

以上