囲碁AI “AlphaGo” はなぜ強いのか?
～ディープラーニング、モンテカルロ木探索、強化学習～

2016/04/15

大槻 知史

目次

- 背景
- 囲碁AIにおけるディープラーニング
- 囲碁AIにおける探索
- 囲碁AIにおける強化学習(など)
- まとめ

AlphaGoに関する最近のニュース

- **AlphaGo以前**
  - 日本の囲碁プラグラムZen等はプロ棋士に4子局で勝利(アマチュア高段者レベル)
  - 人間チャンピオンレベルになるのは10年後位と思われていた
- **2015/10**
  - AlphaGoがヨーロッパチャンピオンのFan Hui 二段に5勝0敗*1
- **2016/01**
  - Google Deep Mind がNature 誌に、AlphaGoに関する論文*1 を発表
    - 市販ソフト(Zen, Crazy Stone …)に対し494勝1敗
- **2016/03/09–15 Google DeepMind Challenge Match*2**

  にて AlphaGo がイ・セドル九段に4勝1敗
  - 日程：3月9日(水)、10日(木)、12日(土)、13日(日)、15日(火)
  - 会場：韓国・ソウル市 フォーシーズンズホテルソウル
  - 賞金：100万ドル (約1億1千万円)
  - 持ち時間：持ち時間2時間／60秒の秒読み3回。

⇒本日は2016/01時点のNature論文を中心にAlphaGoを紹介します

*1: Silver, et al., Mastering the Game of Go with Deep Neural Networks and Tree Search, 2016.
*2: http://www.pandanet.co.jp/event/dmcm

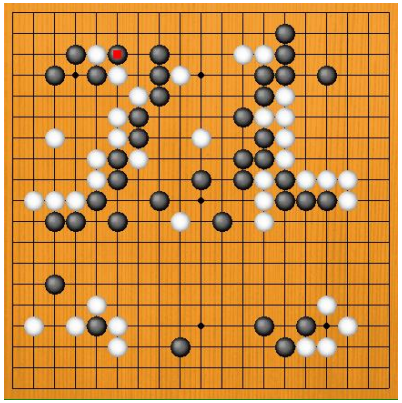
ゲームAI(ゲーム情報学)の進歩

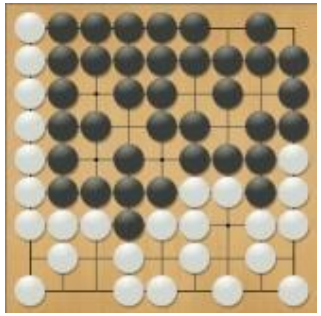
- **オセロ: (探索空間の大きさ*1:～10^{60})**
  - 1997年　Logistelloが世界チャンピオン村上健に勝利
- **チェス: (探索空間の大きさ*1 :～10^{120})**
  - 1997年　IBMのコンピュータであるDeep Blueが世界チャンピオン ガルリ・カスパロフに勝越し
- **将棋: (探索空間の大きさ*1 :～10^{220})**
  - 2013年4月: 第2回将棋電王戦にて、GPS将棋がA級棋士　三浦弘行八段 に勝利
- **囲碁: (探索空間の大きさ*1 :～10^{360})**
  - 2016年3月: AlphaGoが、イ・セドル九段に勝利

*1: 「コンピュータ囲碁の最前線」情報処理学会誌Vol. 54 No.13

囲碁って何?

- **囲碁のルール**
  - 19x19の碁盤の目に、黒番、白番が順番に石を置いていく。
  - **最終的に地が大きい方が勝ち**
  - 相手の石を囲ったら取れる
  - 相手の目には打てない
  - コウ
- **勝ちの判定条件(地の数え方)**
  - 日本ルール
    - 曖昧でコンピュータには扱いにくい
  - 中国ルール(石の数 + 囲んだ地の数)
    - コミ7.5目



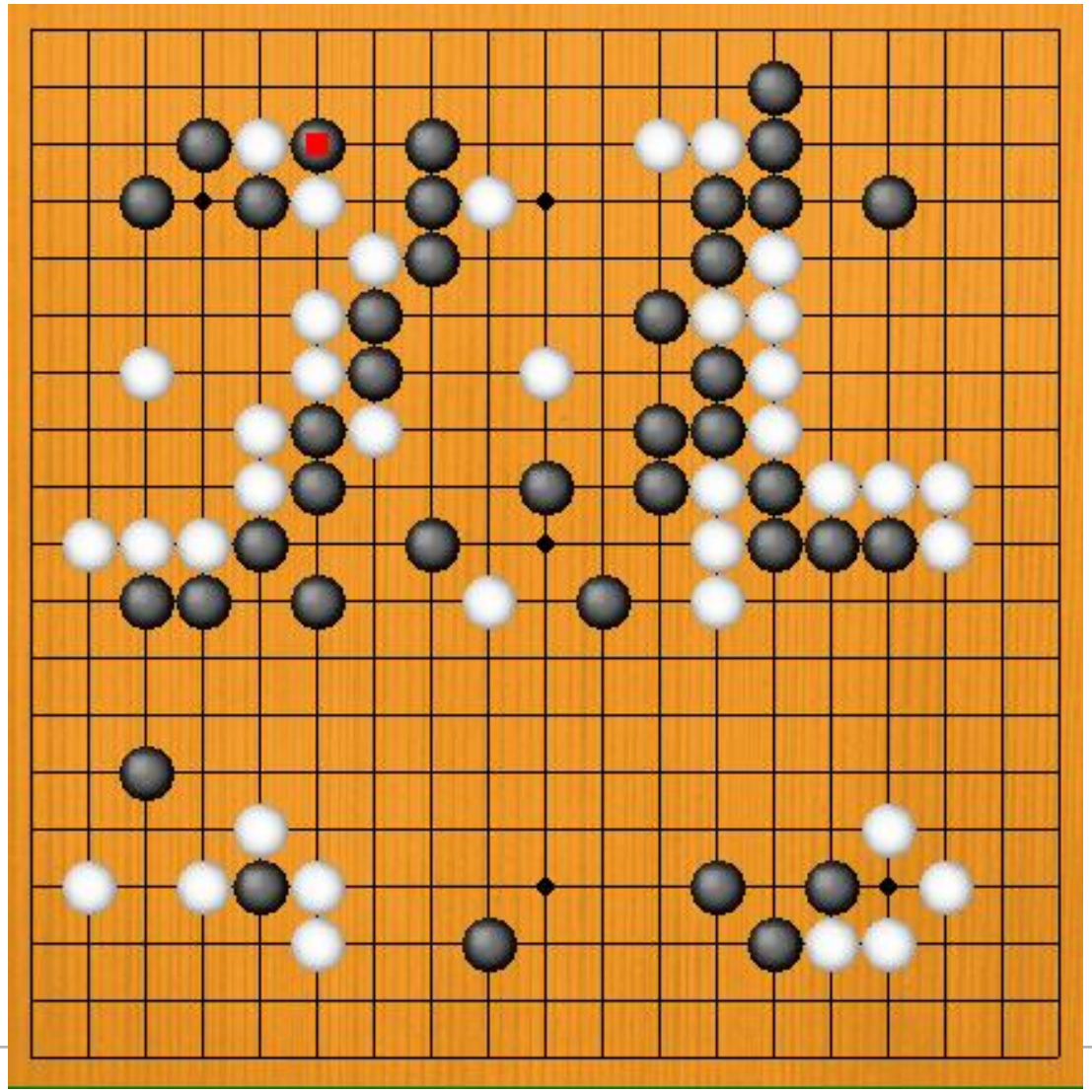


黒地: x目, 白地: y目,
x > y + 7.5 ⇒ 黒勝ち
x < y + 7.5 ⇒ 白勝ち

この場合、x = 45, y = 36
なので黒勝ち

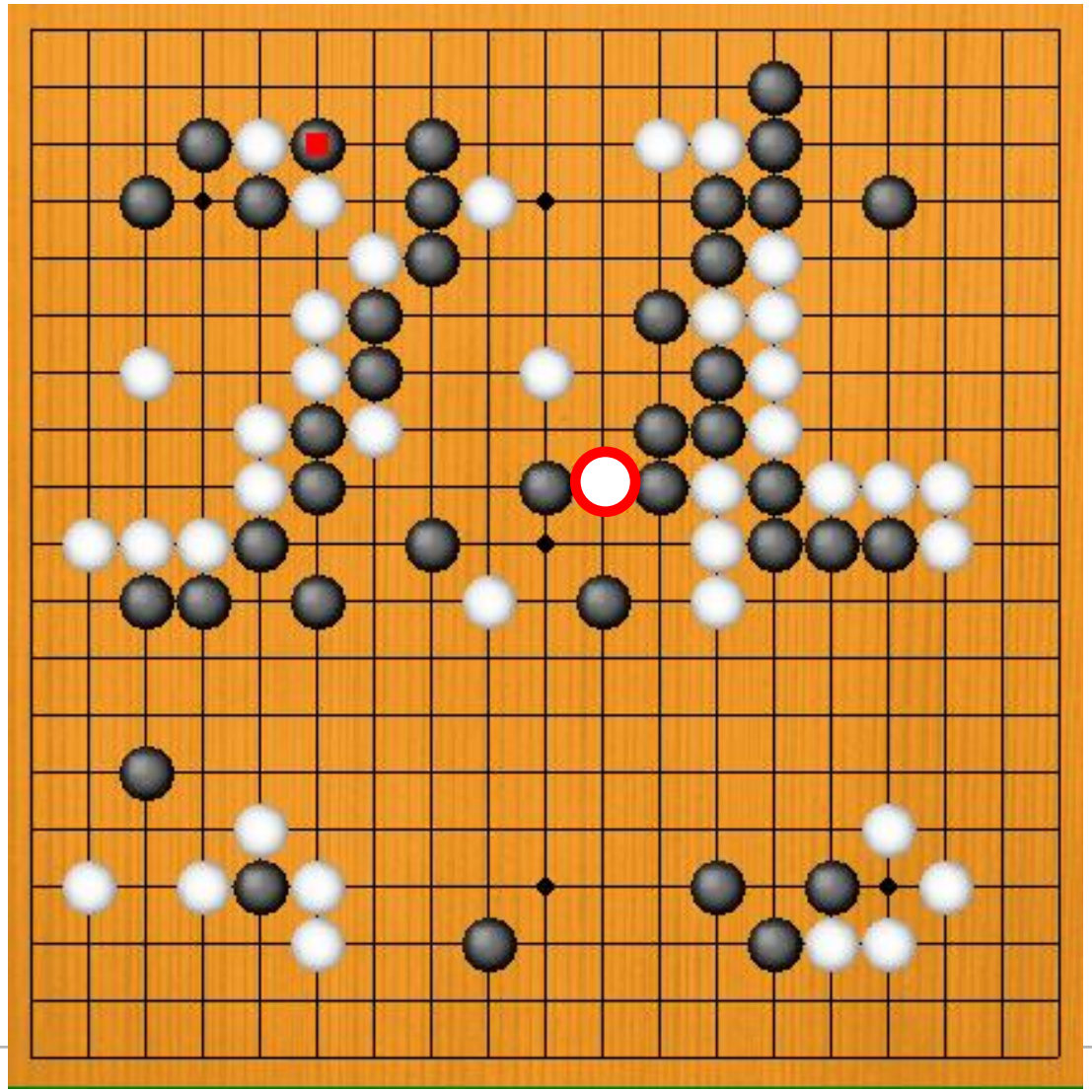
白番(78手目) 次の一手は?

- 白が打てる場所は284箇所、その中からどこに打つか選ぶ



白○が絶妙手!!

- AlphaGoとの対戦で、イ・セドル九段はこの手を発見し勝利!!



目次

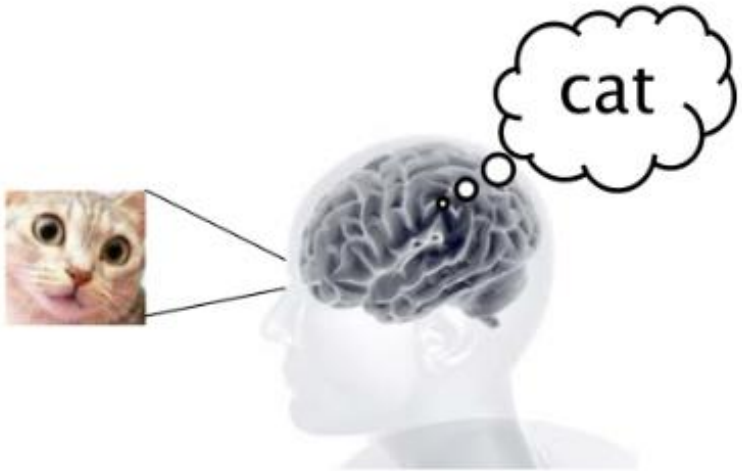
- 背景
- 囲碁AIにおけるディープラーニング
- 囲碁AIにおける探索
- 囲碁AIにおける強化学習(など)
- まとめ

ディープラーニングとは?

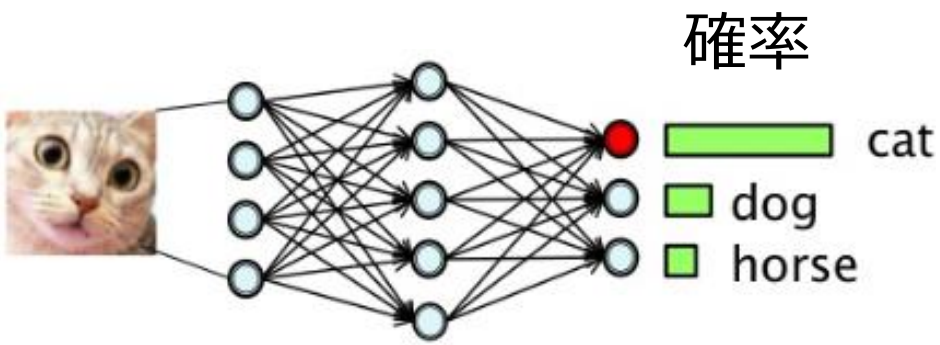
- **(従来よりも深い)ニューラルネットワークによる機械学習の手法**

- **事例: 画像認識(2012年　一般画像認識のコンペで圧勝*2)**

人間の知覚*1



機械学習(ディープラーニング) *1



*1:http://www.slideshare.net/nlab_utokyo/deep-learning-40959442 の図の一部を引用
*2: Krizhevsky, et al, ImageNet classification with deep convolutional neural networks, 2012.

囲碁におけるディープラーニングとは!?

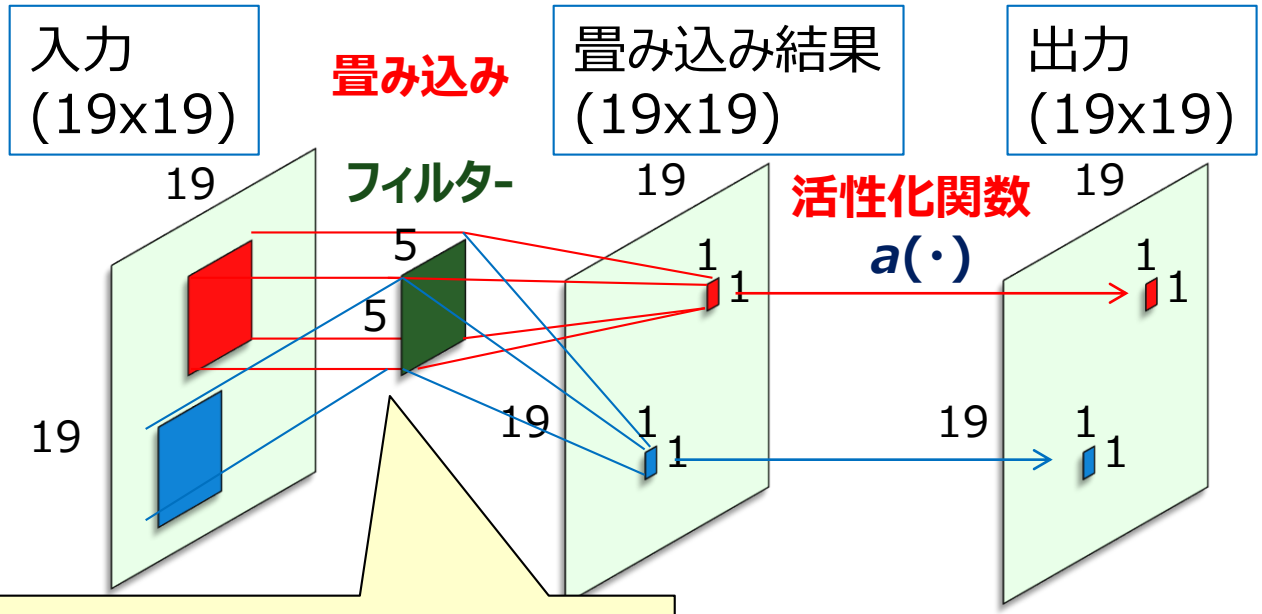
どんな盤面を入力しても、プロみたいな手を打てる何か?

このようなネットワークをプロの棋譜を学習して作りたい!!

畳みこみニューラルネット(CNN)

- **畳み込みニューラルネットとは?**
  - 脳を模倣した、ディープラーニング手法の一種
  - フィルター(受容野に相当)を利用し、**画像(等)の特徴を学習**により自動獲得
  - 深くすることで、局所的な特徴を組合せた複雑な特徴を学習できる

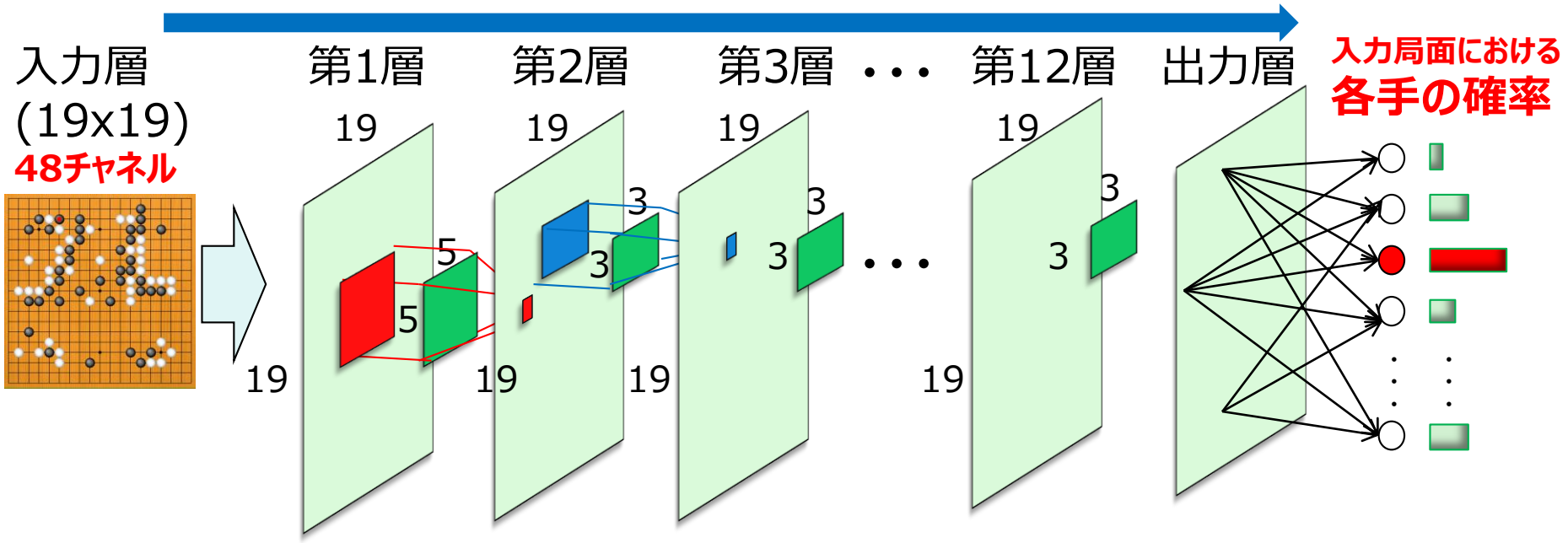
入力19x19 , Filter 5x5 の場合の畳み込みの例



脳とCNNとの共通性
「受容野の局所性」「重み共有」

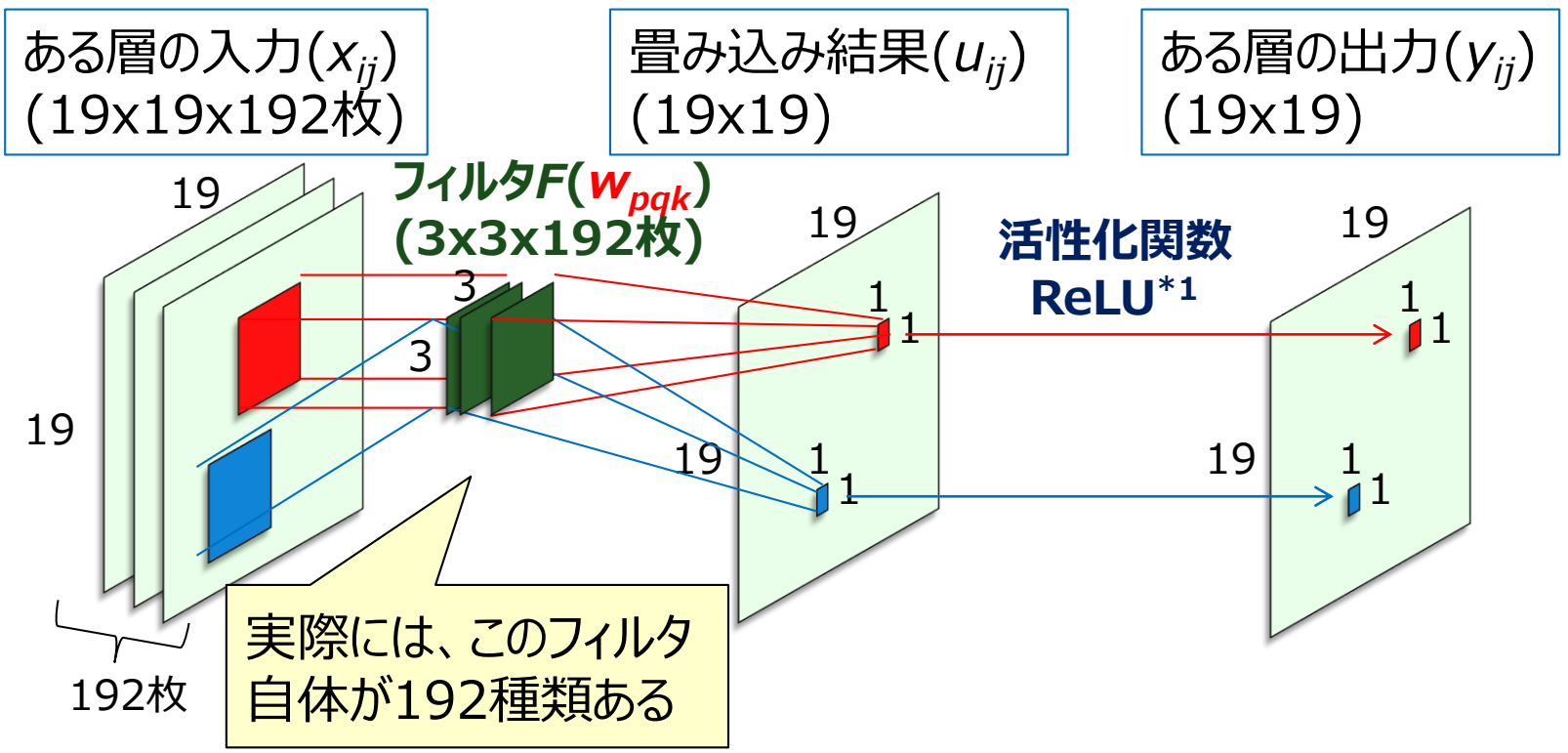
囲碁における畳み込みニューラルネット(CNN)

- **AlphaGoにおけるCNN(SL Policy network)の構成**
  - 入力は48チャネル(黒石/白石の位置、石を取れる位置、シチョウ・・・)
  - 全部で13層[*1]
    - フィルターは各層に192種類ずつあり、1層目のみ5x5, 2～12層は3x3



CNNの構成詳細と計算量

各層は、入力19x19x192枚，Filter 3x3x192枚の畳み込み



畳み込み計算: $u_{ij} = \Sigma_{k=1}^{N} \Sigma_{(p,q)\in F} w_{pqk} \cdot x_{i+p,j+q,k} + b_k$

ReLU 活性化関数: $y_{ij} = \mathrm{MAX}(0, u_{ij})$

パラメータw_{pqk}の個数: 3^2 x (フィルタの種類:192)^2 x(層の数:12) = 約400万個

畳み込みの足し算回数: 19^2 x 3^2 x (フィルタの種類:192)^2 x(層の数:12) = 約14億回

*1: ReLU: Rectified Linear Unit の略

畳み込みニューラルネット(CNN)の学習

- **教師データ: (局面, 強い人間プレイヤの手)の組**
  - インターネット囲碁道場KGSの六～九段の棋譜 16万局(約3000万局面)
- **Back Propagationを利用し、CNNの出力と、人間の手ができるだけ一致するような、フィルタ重みw_{pqk}を求める**
- **1回の畳み込み計算は、GPU(画像処理用プロセッサ)だと　何と4.8 (ミリ秒)[*1] だが、それでも学習には50GPU で3週間!!**

*1: alphaGoの論文による。CPUの場合: 4ギガ(40億回/秒)回計算できるとして、0.36(秒)なので100倍近い!!

AlphaGoのCNN(SL Policy network)の実力

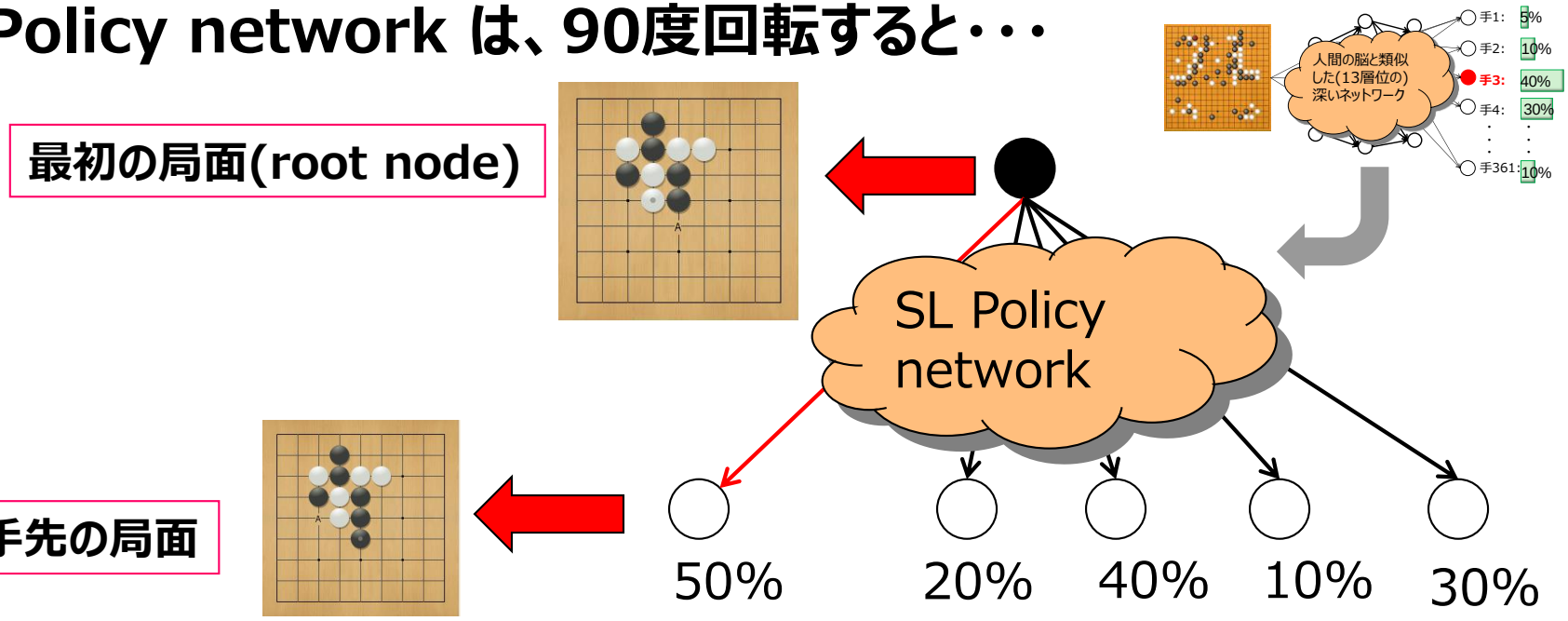
- **局面を与えて一手先の確率を読んでいるだけだが・・・**
  - 従来最高44%程度であった**人間の手との一致率を57%まで高めた**!!
- **囲碁における、CNNの有効性は、他の囲碁AIでも確認**
  - SL Policy network 単体でも、**アマチュア有段者レベル**になる(!?)
  - 従来手法の一部をSL Policy network に置き換えることで、強くなる(!?)

目次

- 背景
- 囲碁AIにおけるディープラーニング
- 囲碁AIにおける探索
- 囲碁AIにおける強化学習(など)
- まとめ

SL-Policy networkから探索へ

- Policy network は、90度回転すると・・・



- この場合

・・・1手進めて・・・勝率を計算して・・・一番勝率の高い手を選ぶことをやっている

- 2手以上進めると、より精度の高い評価が出来そう ⇒ **探索**

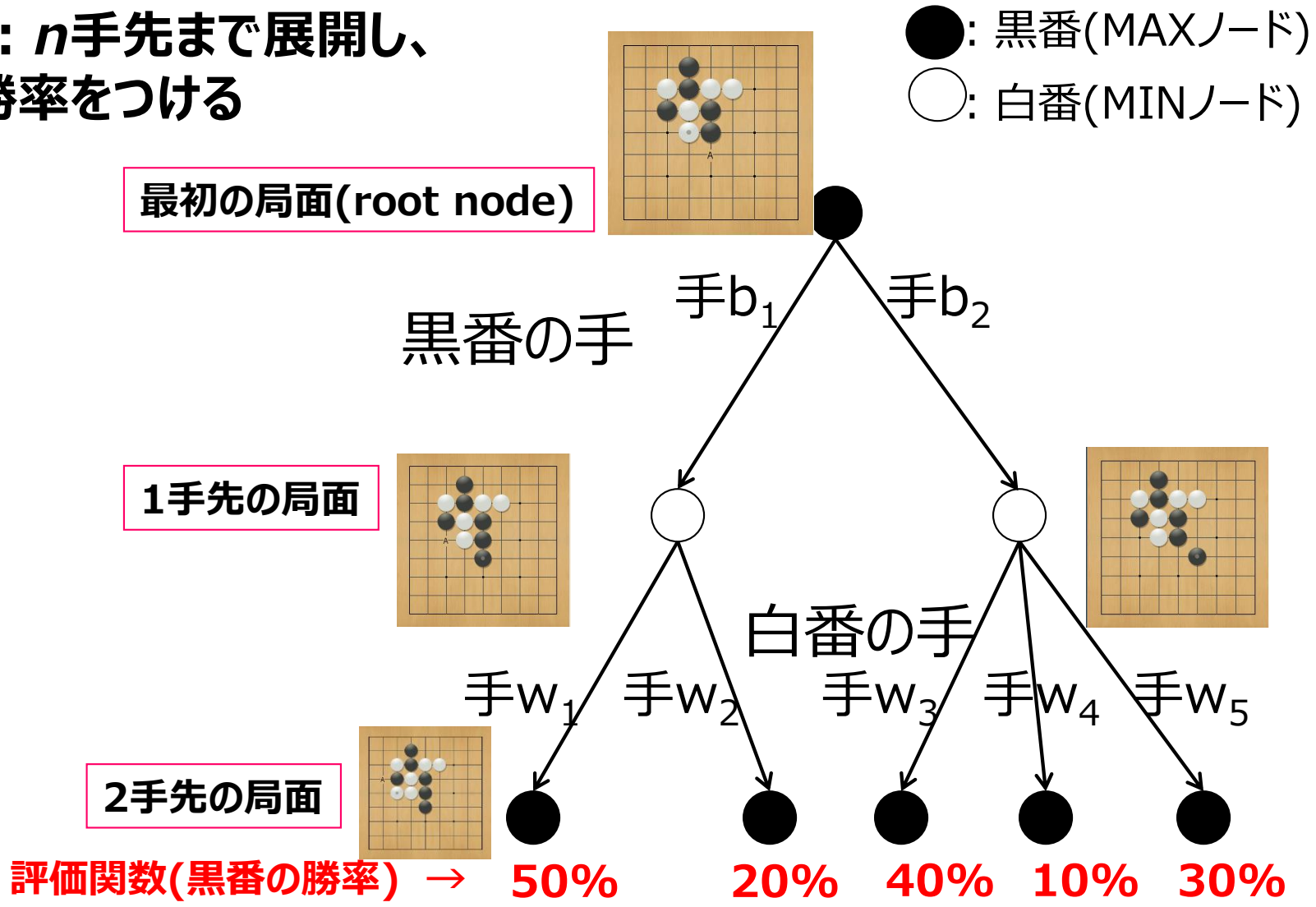
ゲームと探索(囲碁、将棋・・・の比較)

- **従来、将棋・チェスなどのゲームでは、*n* 手先まで手を展開し、*n*手先を評価して、一番良い手順を選ぶ「探索」が主流**
- **一方、囲碁の場合、**
  - 精度の高い評価関数を作るのは難しい
    - 石のつながりや強さ、死活などを数値化することが難しい
  - 候補手の数が非常に多く、深く探索することは難しい(初期局面は361手)

	将棋	囲碁
評価項目	駒の数、駒の位置	石のつながり 死活
評価関数の設計	駒の価値:明確 駒の位置: 王との相対位置で大体OK	ちょっとした形の違いで、「つながり」や死活が変わるため困難
候補手の数	初期局面:30	初期局面:361
探索アルゴリズム	**しらみ潰し探索ベース**	**モンテカルロ木探索ベース**

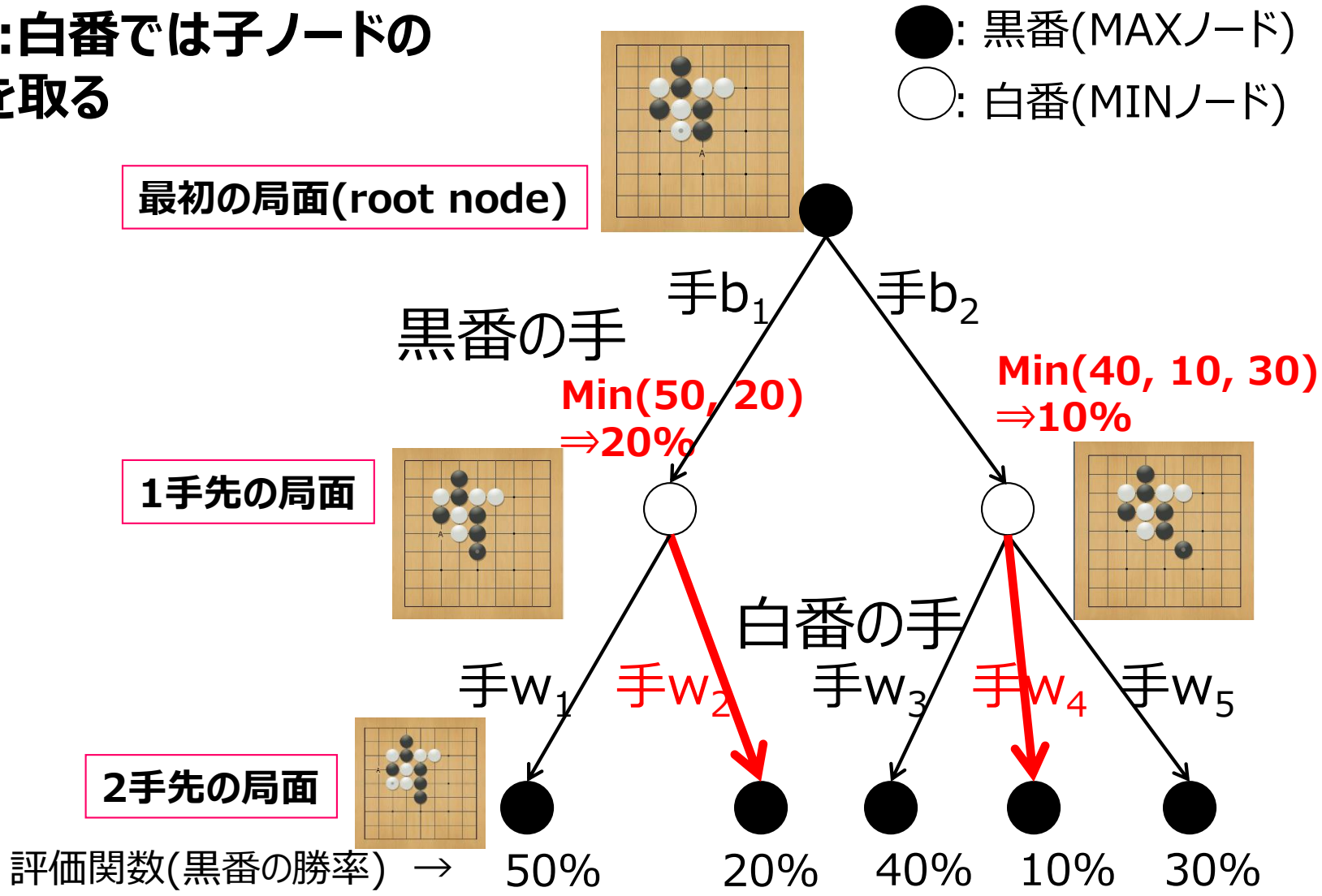
従来のゲーム木(Min/Max 木)探索

Step 1: *n*手先まで展開し、末端に勝率をつける



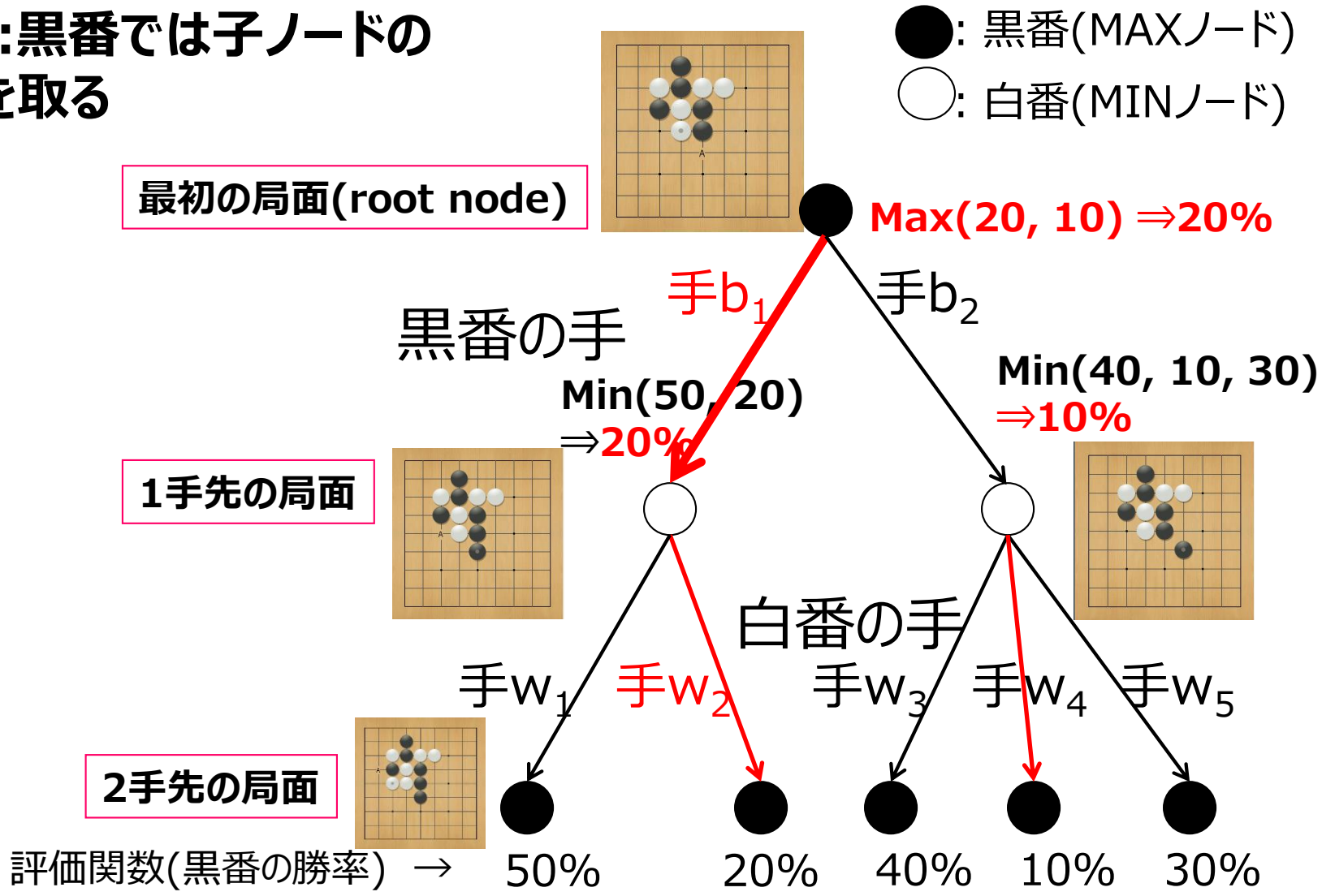
従来のゲーム木(Min/Max 木)探索

Step 2:白番では子ノードの最小値を取る



従来のゲーム木(Min/Max 木)探索

Step 3:黒番では子ノードの最大値を取る



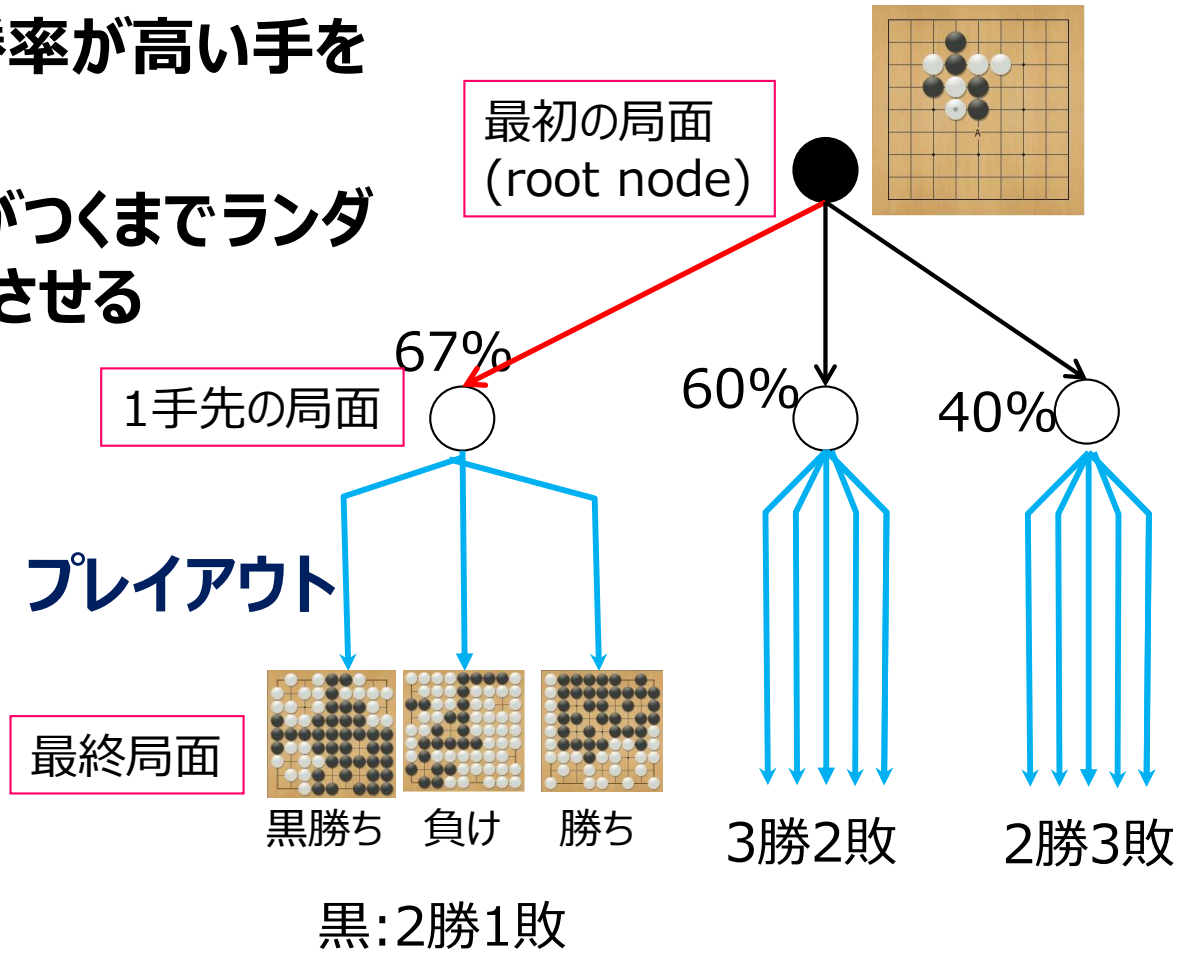
ゲーム木探索のポイント: 枝刈り ＆ 評価関数

・将棋等の「しらみ潰し」探索では、枝刈りと評価関数が重要
  ・**枝刈り/深さ延長**: いかに重要な変化を深く読むか
  ・**評価関数**: 末端局面をいかに正確に評価するか

・一方、囲碁では、候補手が多く正確な評価が困難なため、「しらみ潰し」は難しい

原始モンテカルロ

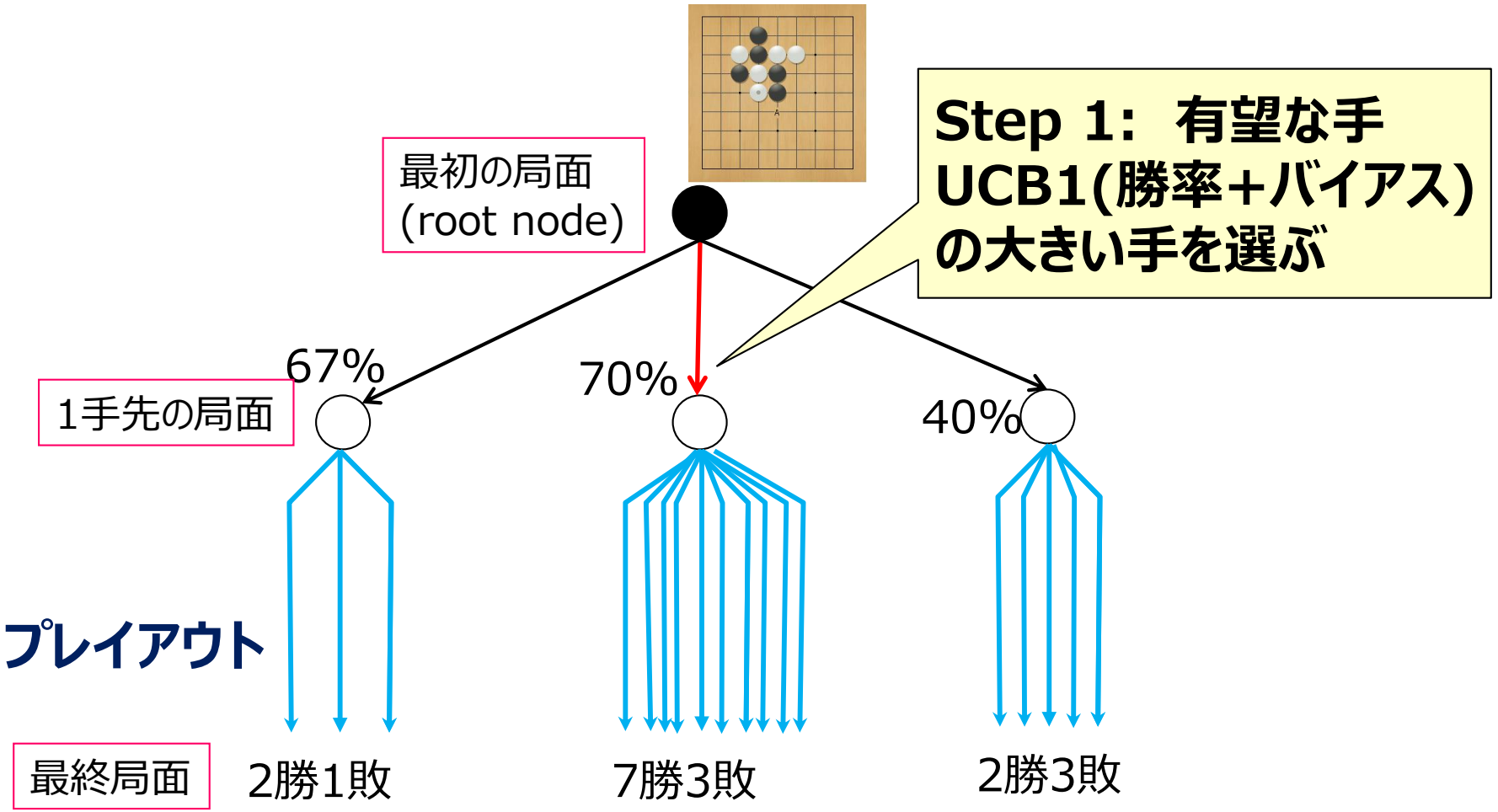
- 1手進めたところからランダムにプレイアウトして、最も勝率が高い手を選ぶ
- プレイアウト: 決着がつくまでランダムに手を打って対戦させる



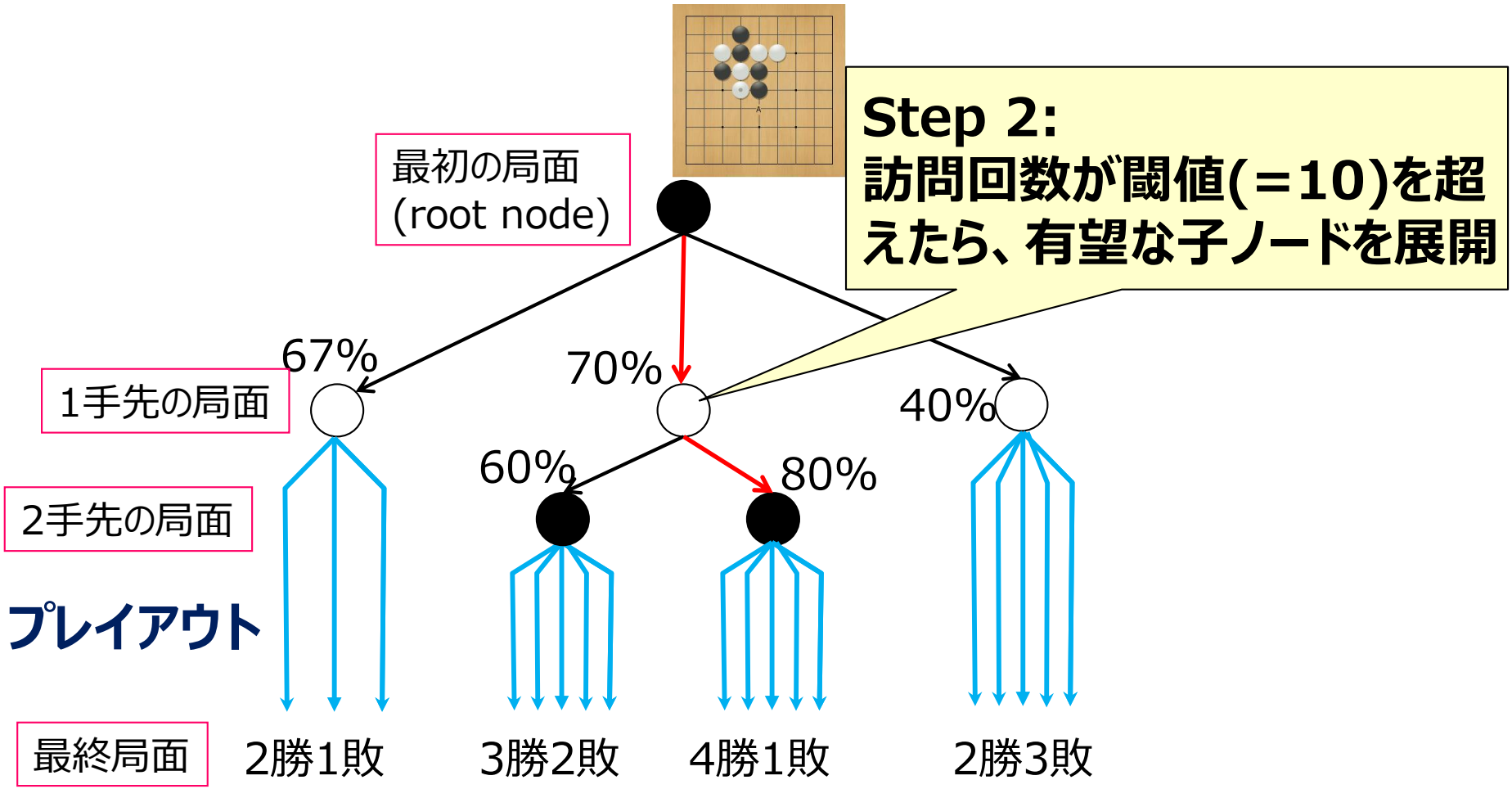
原始モンテカルロの問題点

- **原始モンテカルロは**
  - 一番勝率の高い手を選ぶので、一見これでよさそう。
  - しかし・・・プレイアウトの中では、ランダムに探索するため、相手に一つだけ良い手があるような場合、判断を誤るリスク
- **そこで、「有望な手」に対しては、より深く調べて、本当に良い手かどうかを確認する⇒「モンテカルロ木探索」**

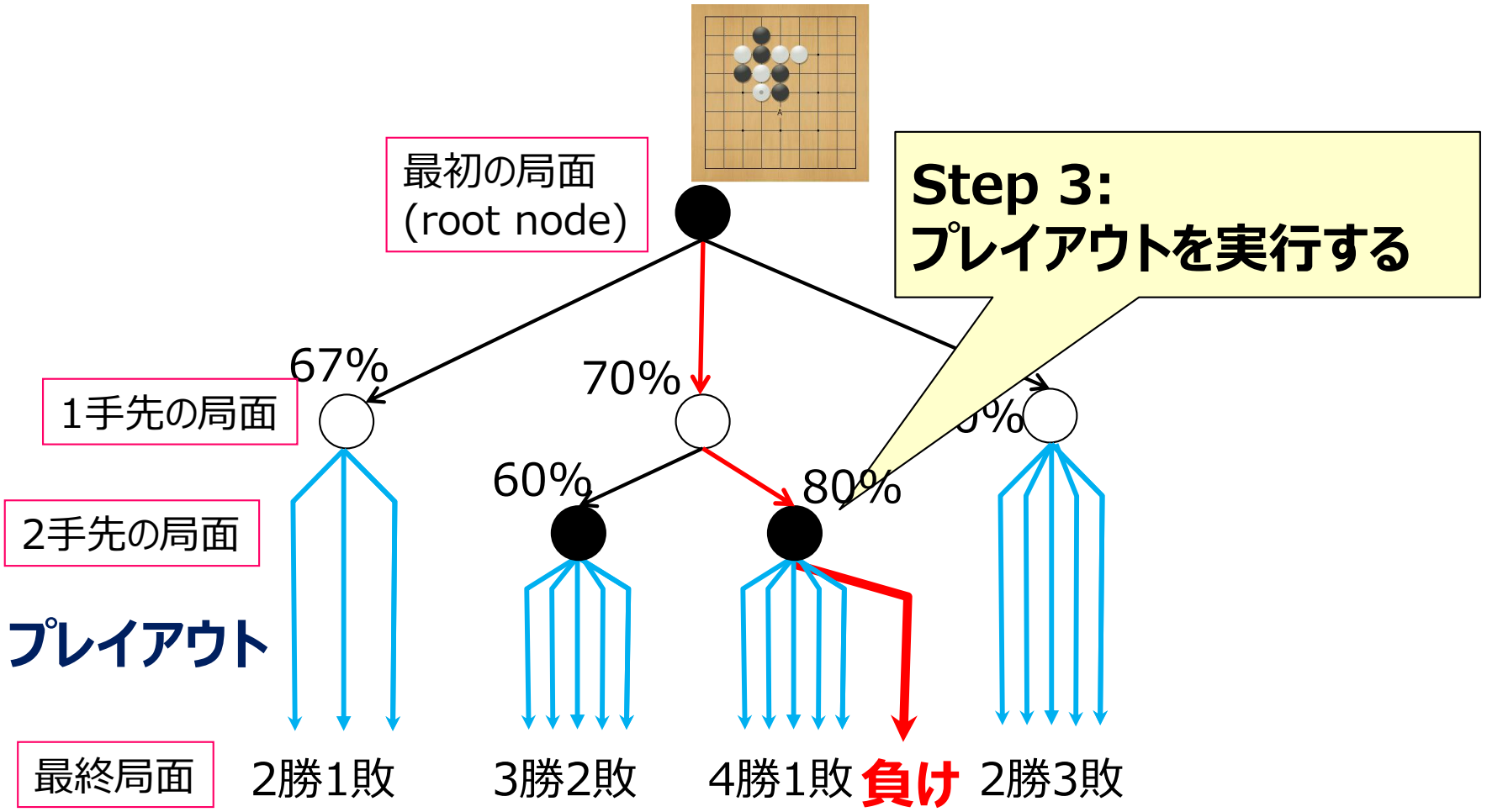
モンテカルロ木探索(Step 1: Selection)



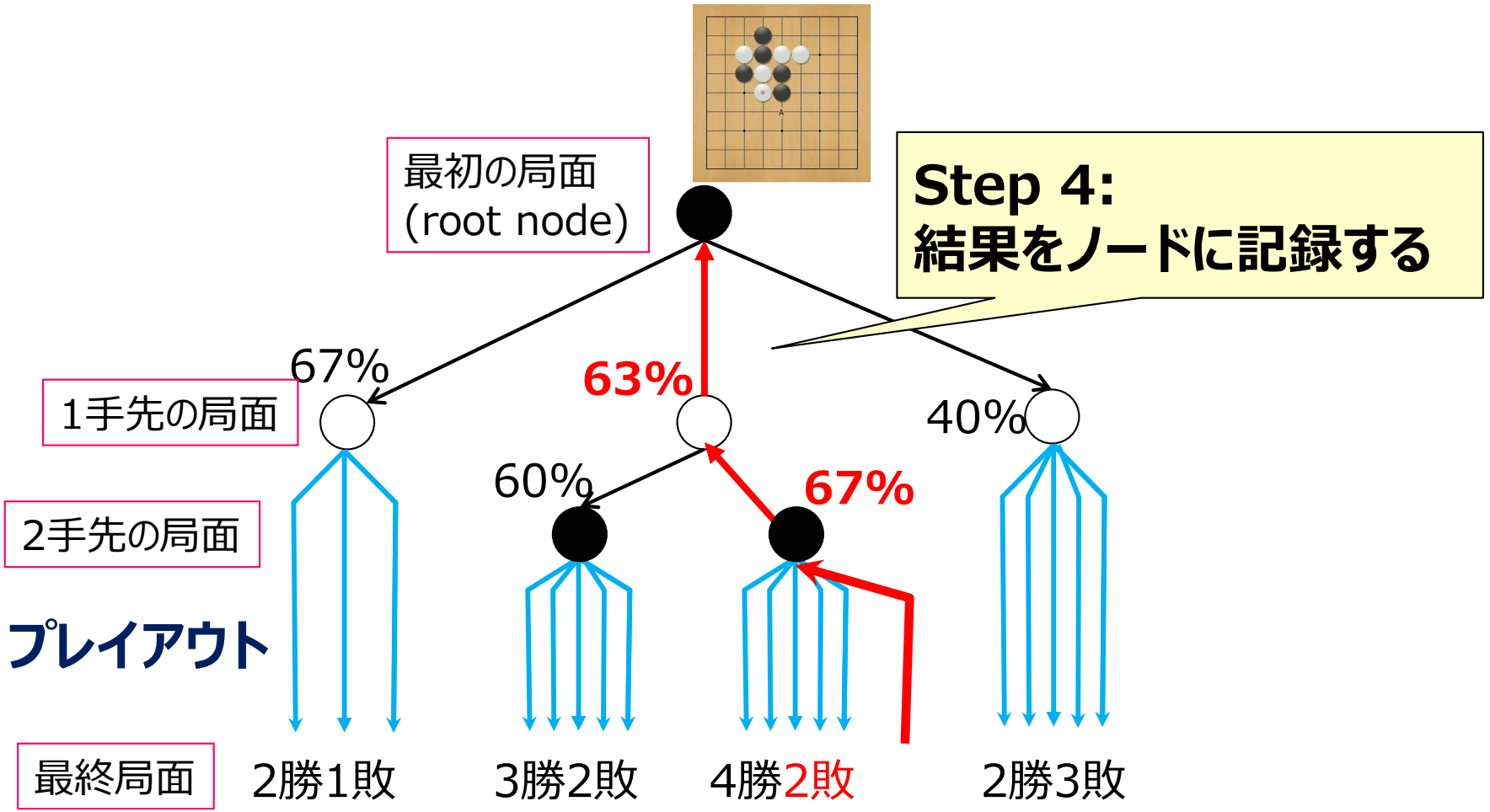
モンテカルロ木探索(Step 2: Expansion)



モンテカルロ木探索(Step 3: Evaluation)



モンテカルロ木探索(Step 4: Backup)



SELECTION のポイント: UCB1を使う

- **有望な手の選び方(UCB1[*1])**
  - 基本は勝率、でも、たまたま一回失敗しただけで選ばれなくなるのは避けたい
  - そこで、**UCB1(勝率とバイアスの和)**が最大になるようにする
  - 多腕バンデット問題に対する一般的な手法

勝率: この局面以下の勝率

バイアス:探索回数が少ない場合に大きくなる

$$UCB1 = (w/n) + (2\log t/n)^{1/2}$$

n: この局面の総プレイアウト回数
w: この局面の勝ち数
t: 兄弟ノード全体の総プレイアウト回数

- **UCB1を使う場合、最善手のプレイアウトの回数が最大となるがことが、理論的に保証される[*2]**

*1:Auer et al, Finite-time analysis of the multi-armed bandit problem, 2002.
*2:十分試行回数が大きい場合

モンテカルロ木探索の性質とその改良

- **従来の優劣評価(≒勝ちの大きさ)ではなく、勝率(勝ち数)で評価**
  - 「優勢な時は最善でなくても確実な手」、「不利な時は勝負手」を打つ
- **プレイアウト数を増やすほど強くなる:**
  - AlphaGoの場合、初期局面で1000(回/秒)プレイアウトを試行できる
  - 並列化が重要
- **プレイアウトの質を高めるほど強くなる:**
  - プレイアウトの速度低下を抑えながら、(ランダムより)手の質を上げる工夫がある
  - CNNによる方法は正確だが、1局面の評価に4.8(msec)もかかるので、プレイアウトには使えない
- **木をうまく展開するほど強くなる:**
  - RAVE, Progressive Widening …(でもAlphaGoは使っていないらしい)

目次

- 背景
- 囲碁AIにおけるディープラーニング
- 囲碁AIにおける探索
- 囲碁AIにおける強化学習(など)
- まとめ

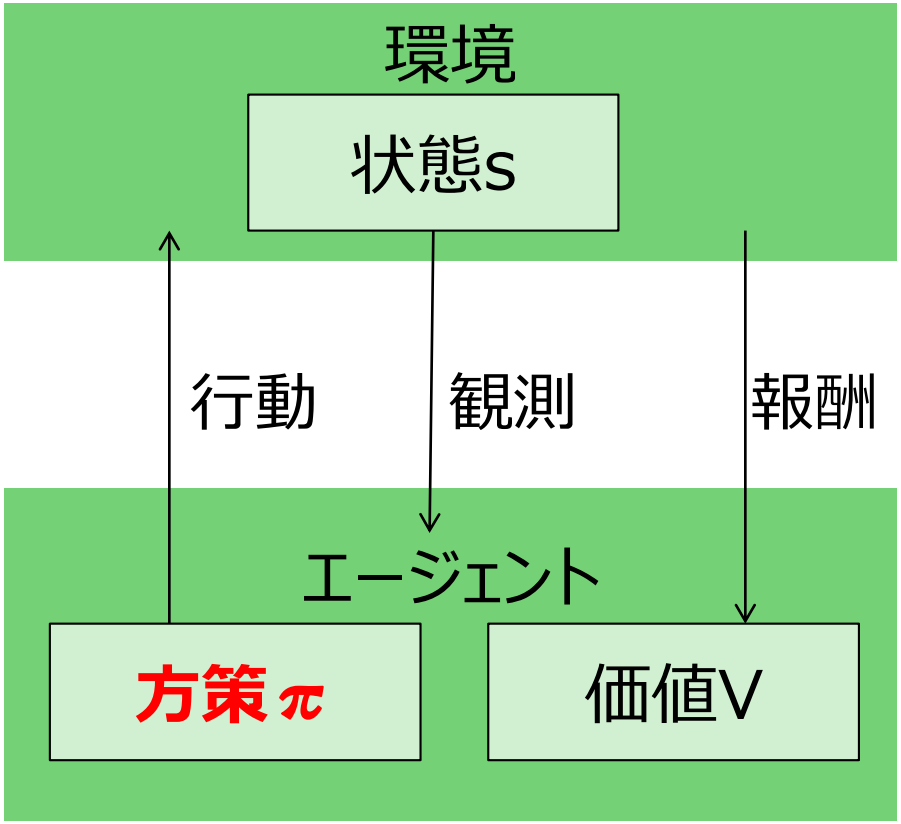
Policy networkと、モンテカルロ木探索を融合したい

- これまで以下の2つを説明
  - SL-Policy network(CNN): 高い人間プレイヤとの一致率
  - モンテカルロ木探索: UCB1によるSELECTIONがポイント
- これらを融合したいが、その前に、CNNをモンテカルロ木探索用にチューニングしたい

  (一般に、深さ1で有効なCNNの重みと、モンテカルロ木探索で有効な重みは異なるはず・・・)

強化学習とは?

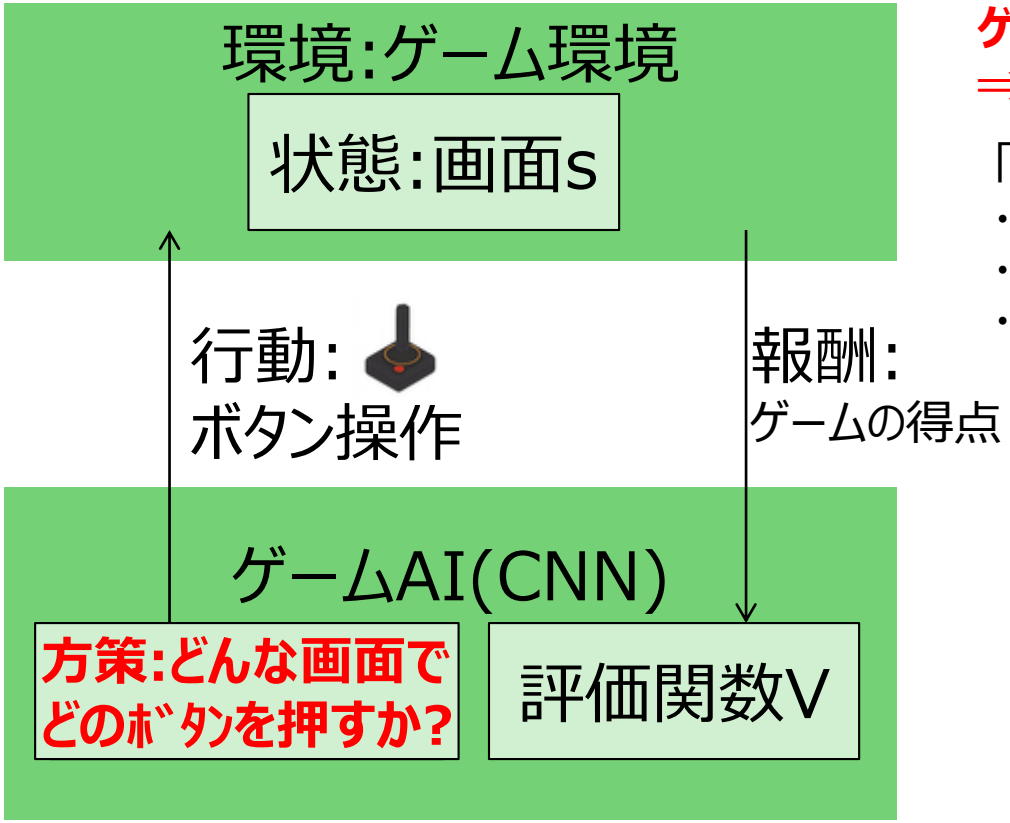
- **強化学習:未知の環境の中を探索しながら期待報酬和を最大化するためのエージェントの行動原理*1**
  - 正解は与えられないが選んだ答えの「良さ」(報酬)を元に、行動原理(方策)を改善



*1: https://www.sat.t.u-tokyo.ac.jp/~mak/20140319-makino.pdf より引用

強化学習の例: ゲーム操作を学習(DQN)

- **Deep Q Learning Network (DQN) *1: CNNと強化学習により自動的にゲーム操作を学習⇒人間を超えるゲームAI**



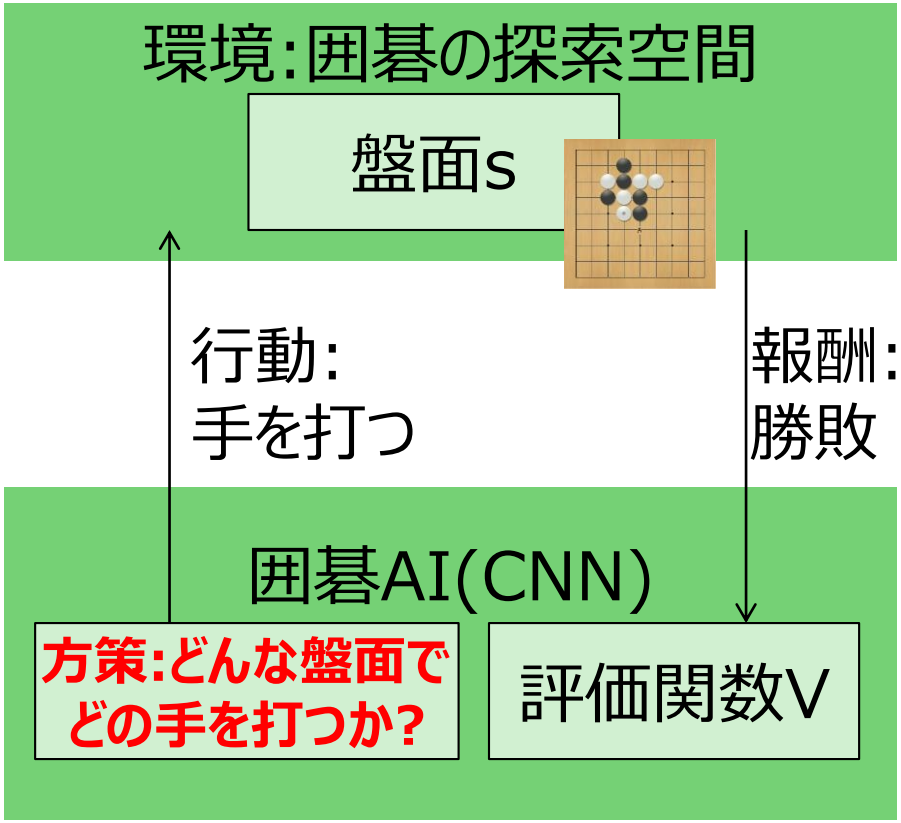
ゲームAIに、ゲームをプレイさせながら学習
⇒最初はランダムだが、どんどん上達

「ブロック崩し」ゲームの場合
- 最初はよくボールを落とす
- 徐々にボールを落とさなくなる
- 壁に穴を開け、ボールを壁の上に転がす

*1:: Mnih, et al. Human-level control through deep reinforcement learning, 2015.

囲碁における強化学習とは?

- **囲碁の場合は、自己対戦しながら、パラメータをチューニング**



囲碁AIに、実際に
自己対戦させることで
学習を進められる!?

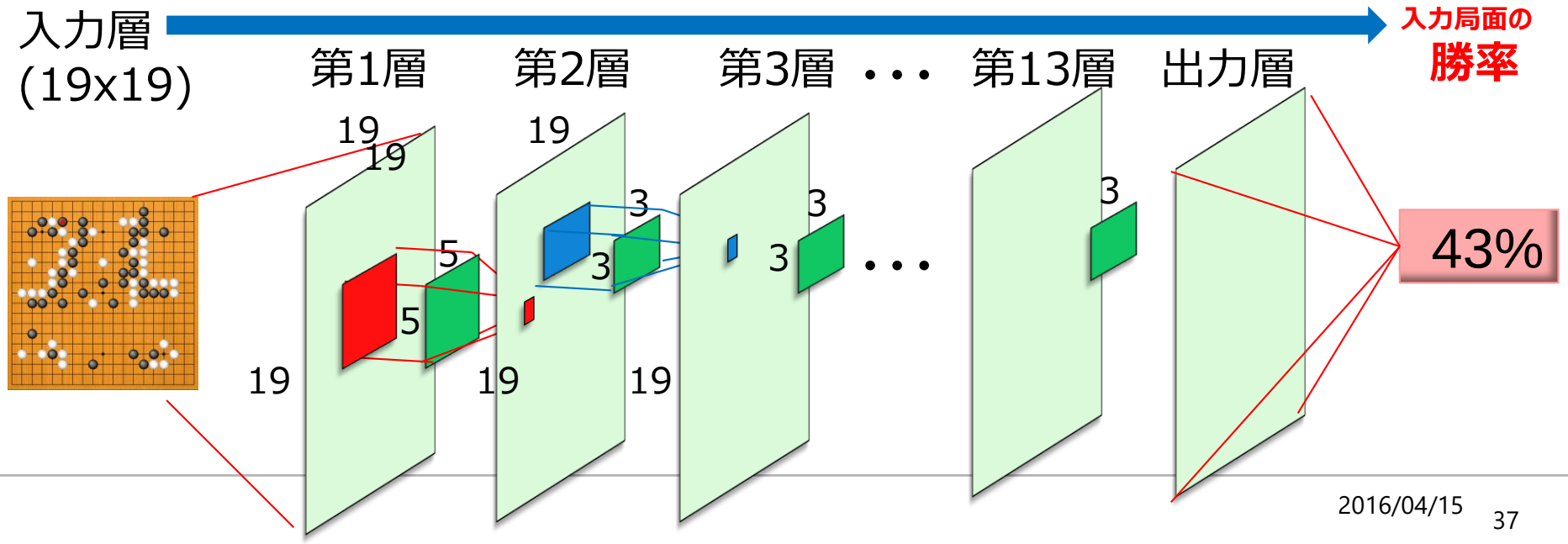
強化学習によりCNNを洗練(RL-policy network)

- **SL-policy network を初期値とし、「ゲームの勝利」を報酬として、自己対戦により学習する ⇒ RL-policy network**
  - 勝った時は、勝つに至る手を出来るだけ選ぶように
  - 負けた時は、負けるに至る手を出来るだけ避けるように
- **今度は、自己対戦により、ゲームの結果を得る必要があり、やはり膨大な計算量が必要**
  - 50GPUで約1日かかる

ある方策を元に手を打ってから、勝ち負けが決まるまでには手数がかかる
(⇔ ブロック崩し: あるポリシーで動かした後、のボールが落ちてくるには時間がかかる)

局面の勝率を評価できるValue Network

- 局面を入力とし、勝率を出力するCNN
- 教師データ: (局面, 最終的な勝敗)との組合せ
  - RL-Policy network を用いて自動生成
- CNNによる勝率が、実際の勝敗に近づくようにFilter重みを更新
- 3000万局面のデータを利用し、50GPUで1週間かかる
- 従来困難とされた囲碁の評価関数(Value Network)ができた!!



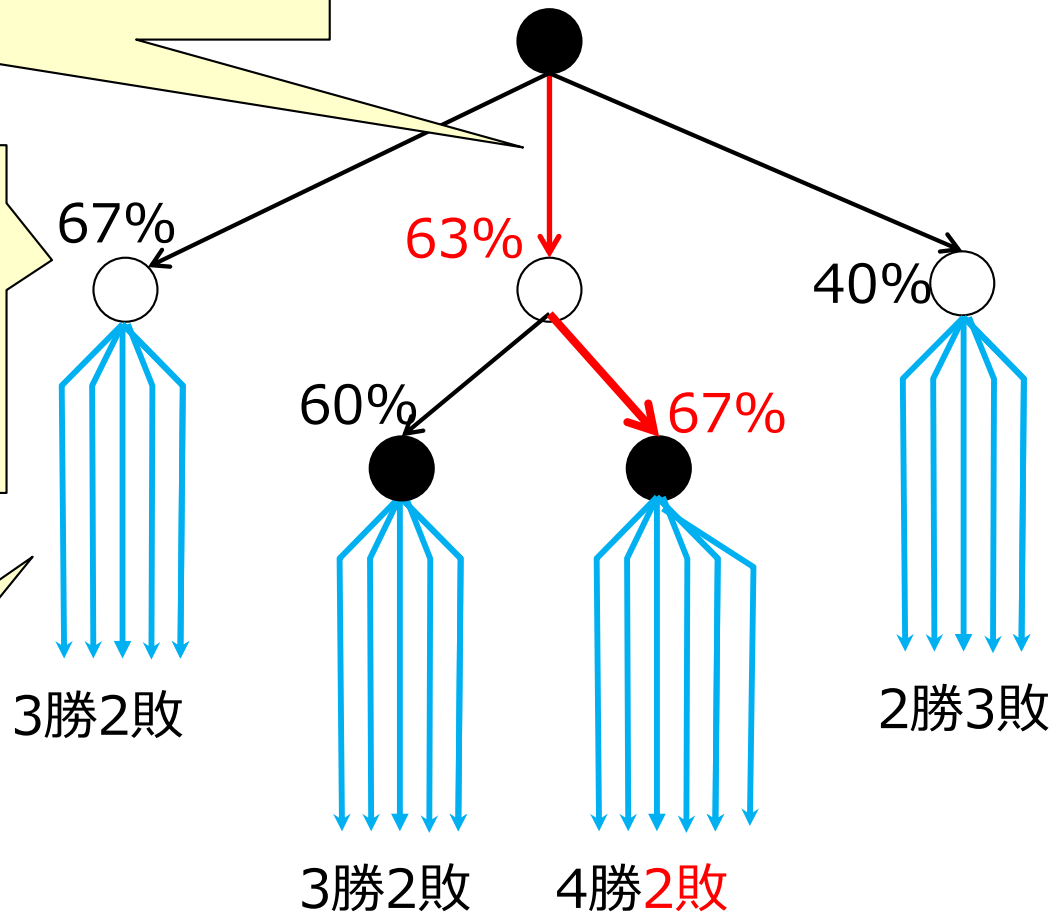
CNN + モンテカルロ木探索 = AlphaGoができた!!

ポイント1:バイアス評価の改善
バイアス + Policy networkによる手の事前確率
⇒ありそうな手順をより深く展開できる

ポイント2:勝率評価の改善:
プレイアウトの勝率 +
Value networkによる勝率評価
⇒プレイアウトを補完する評価が可能

ポイント3:大量のGPU・CPU
⇒CNN評価(GPU 176個)
プレイアウト(CPU 1202個)
の高速化

従来のモンテカルロ探索:
UCB1 = 勝率 + バイアス
の高い手を選択



目次

- 背景
- 囲碁AIにおけるディープラーニング
- 囲碁AIにおける探索
- 囲碁AIにおける強化学習(など)
- まとめ

Alpha-Goはなぜ強いのか?

- **従来手法でも既に結構強かった(アマチュア有段者レベル)**
- **CNNにより、良い手を深く探索できるように(SL Policy network)**
  - (強い人間プレイヤの手との一致率: 44%⇒**57%**)
- **Policy network を徹底的に強化学習(RL Policy network)**
- **CNN評価関数(Value network)により局面評価精度向上**
  - これまで誰も成功していなかった偉業!!
- **以上を融合したモンテカルロ探索を大量のGPU/CPUにより高速化**
  - 莫大な計算量が必要なるプロセスをやりきったところが凄い!!

まとめと今後の課題

- **AlphaGo のNature論文の内容を解説**
  - 従来のモンテカルロ木探索とCNNを組合せることで実力が向上
    - 囲碁では、まだ人間チャンピオンが勝つ余地があるかもしれないが、いずれ勝てなくなるだろう
- **最近は他の分野でも・・・**
  - 画像認識/音声認識/機械翻訳　など他の多くの分野でも、従来モデルを、CNN等に置き換え、性能が飛躍的に向上
  - 10,20年後に日本人の職業の49%は奪われる!?[*1]
- **ゲーム情報学の研究者も、囲碁が終わることで、目標喪失!?**
  - 正解が不明確な問題、不確定な問題はまだまだ苦手
    - 人が強くなるのをサポート(教育モード)
    - わざと負けて人を喜ばせる(接待モード)
    - 不完全情報ゲーム(ブリッジ、麻雀・・・)

*1: NRIプレスリリース　http://www.nri.com/Home/jp/news/2015/151202_1.aspx

参考文献

- **AlphaGoに関して**
  - Silver, et al., Mastering the Game of Go with Deep Neural Networks and Tree Search, Nature, 2016.
  - 伊藤毅志 , 村松正和,「ディープラーニングを用いたコンピュータ囲碁～ Alpha Go の技術と展望～」, 情報処理学会誌4月号, 2016.
- **コンピュータ囲碁に関して**
  - 美添一樹, 山下宏 , 松原仁編,「コンピュータ囲碁 ―モンテカルロ法の理論と実践」, 2012
- **畳み込みニューラルネット(CNN)に関して**
  - 麻生英樹, 安田宗樹, 前田新一, 岡野原大輔 ,「深層学習 Deep Learning (監修:人工知能学会) 」, 2015.
- **強化学習に関して**
  - 牧野貴樹,「強化学習をベイズで理解する」, 2014. https://www.sat.t.u-tokyo.ac.jp/~mak/20140319-makino.pdf

ご清聴ありがとうございました

２０１６/７追記

- **AlphaGo(Nature論文)のSL-Policy Network を再現してみました*1。普通に囲碁ソフトとして遊べますので、興味のある方は下記をご覧ください。**

  http://home.q00.itscom.net/otsuki/delta.html

- **GoGuiなどの囲碁対戦用GUIを使えば、普通のパソコンで簡単に対戦できます!!**

*1: 192filterで強い人との手の一致率が54%程度に達しました。
（Nature論文の55%に近い値です）

２０１７/７,2017/11追記

- 本資料を基に、アルファ碁の技術を著作

『最強囲碁AI アルファ碁解体新書』

にまとめました(2017/7)

- さらに上記から、アルファ碁ゼロの技術部分の差分をpdfにまとめました(2017/11)

こちらもよろしくお願いいたします!!