ブルータスAI
~Brutus AI~

2017年AlphaGo Zeroという囲碁AIが作られ，囲碁チャンピオンをも倒してしまいました．さらに，このAIは人間が囲碁の戦略を教えること，また人間同士の対戦の棋譜を与えることは全くせずに，ルールだけから学習したということから，注目を集めました．その後，そのアルゴリズムをチェスや将棋などの一般のゲームに適用できるようにした，AlphaZeroが登場し，こちらも好成績を残しています．

この展示では，同じAlphaZeroを用いて，ブルータスというボードゲームを学習させました．展示では，実際のAIと対戦することができます．

あなたはこのAIに勝てますか？

理論

私たち人間がブルータスや将棋のようなボードゲームで遊ぶとき、最善な手を打つためにいくつか先の盤面を予測する「先読み」をすることがあるのではないでしょうか．ここで使われているAIも，この「先読み」がうまくできるように学習していきます．

強化学習とは

はじめのうちはAIはどの手がうまい手なのかはわかりません．そこで，AI同士の対戦を繰り返し，学習データを集めていきます．そして，勝利につながった有力な手を多く打てるように学習していきます．このように最適な行動(=良い手)が未知であるけれども，そこから得られる結果(=ゲームの勝敗)が利用できる場合に，最適な行動を学習する手法を強化学習といいます．

先読みするには

学習が進むと盤面の情報を入力とすることで，次の一手としてどの手が有力であるかと，この盤面がどの程度良い盤面なのかを予測することができるようになります．強化学習によって予測された結果に従って，有望な手ほど深く先読みします．そして，その先読みの結果をもとに次の一手を決定します．ここで，先読みには，モンテカルロ木探索という方法が用いられています．

具体的な学習方法について

説明で出てくる添え字 a は次の一手を表していると考えてください.

ここでゲームのAIを関数 $f_\theta(s)$ として表しておきます．引数 s は盤面の状態で，θ はパラメータです． $f_\theta(s)$ はそのときの勝ちそうな手番を表す変数 v (1に近いほど先手勝利，−1に近いほど後手勝利)と次に打つべき手の重み $\boldsymbol{p}=(p_a)$ の組を返すように学習していきます．すなわち，

$$f_\theta(s)=(v,\boldsymbol{p})$$

です.

Alpha Zeroは，まず自分どうしを対戦させます．そして盤面の状態sとそのときに次の一手についてモンテカルロ木探索で訪れた回数 $\boldsymbol{n}=(n_a)$ とその勝負に勝ったか負けたかを表す変数 $z=\pm1$ を記録していきます.

$\boldsymbol{\pi}=\boldsymbol{n}^{1/t}$ (t は温度と呼ばれる人間が調整するハイパーパラメータです)とします.この手法ではある関数を最小化するように学習を進めていくのですが，その関数をLとすると

$$L=(v-z)^2-\boldsymbol{\pi}^{\mathrm{T}}\log\boldsymbol{p}+\lambda\|\theta\|_2^2$$

となります．棋譜のランダムサンプリングしながら， L が小さくなるようにパラメータ θ を変化させていきます.

第1項は v が z に近くなるように学習させています.

また，第2項はややわかりにくいですが，$\boldsymbol{\pi}$ とlog $\boldsymbol{p}$ の内積なので，ベクトルとして $\boldsymbol{\pi}$ と log $\boldsymbol{p}$ ができるだけ同じ方向を向くように学習させています(ともに非負ベクトルです)． $\boldsymbol{\pi}$ はモンテカルロ木探索で訪れた回数を反映させたベクトルとなっているので， log $\boldsymbol{p}$ もそのような値に近づいていきます.

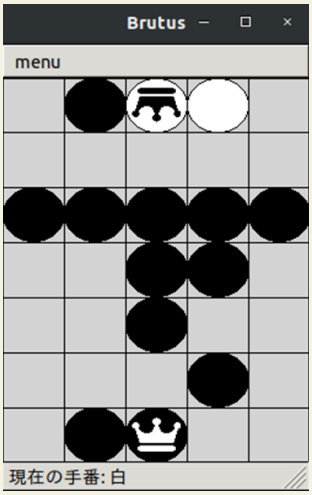
第3項は正則化項と呼ばれ，モデルの過学習を防ぐために導入されます．その強さをハイパーパラメータ λ で調整します.

モンテカルロ木探索(MCTS)について

最後に，「先読み」をするために必要不可欠なモンテカルロ木探索(MCTS)について説明していきます．MCTSでは学習させた関数 $f_\theta(s)$ を用いて手の先読みを行います.

MCTSは何手か先の盤面とその盤面がどれだけ優れているかという評価点をの構造を木と呼ばれるグラフ構造で持つのですが，選択・展開・評価・バックアップという手法を繰り返しながら，その評価点を更新していくことにより，次の手を決めています.

AlphaZeroの手法では通常のMCTSと選択・展開・評価の仕方が異なり，ニューラルネットワークの出力 $(v,\boldsymbol{p})$ をうまく利用し，有効な手を重点的に探索し効率的かつ高精度に先読みを行っています.



参考文献

[1] Silver, D. et el. Mastering the game of Go without human knowledge (2017)

[2] Silver, D. et el. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm (2017)