

# no-policy eg

## 概要

2024年版egから変わりはありません。

ポリシーネットワークで自己対局の探索数分布を学習する代わりに、自己対局中の手の評価を学習しました。

## 現状と課題

CrazyStoneでは、パターンと手作り特徴量を用いた着手のレーティングでの前方枝狩りが導入されました。

Darkforest, AlphaGoなどは、ポリシーネットワークと呼ばれるDCNNで強いプレイヤーの棋譜を学習し着手を予想するようになり、飛躍的に強くなりました。

その後、Alpha Go ZeroやKataGoでは学習対象が、ある局面での1手ではなく、MCTSの探索数分布に拡張されました。

AlphaGoやそれ以降の成功が示すようにポリシーネットワークは有効な仕組みですが、いくつかの課題があります。

1. 人間にとって、着手確率という概念になじみがない。  
90%の手や0.3%の手とはあまり言わない。(囲碁AIの普及で最近はそうでもないかもしれませんが。)
2. 強化学習の際に同等の価値の手が複数ありポリシーの出力に偏りがある場合、そのモデルを使った自己対局での探索回数の分布にも事前確率と同様の偏りが発生する。  
これは、対局者としてのAIにとっては効率的に探索木の枝を絞り込めるため有用な特性である反面、人間が検討や学習に用いる場合、理想的な手と一致しているとは限らない。
3. 自己対局でポリシーネットワークの教師データとする局面では十分なプレイアウト数が必要になる。  
Gumbel MuZeroで示されているように囲碁では200プレイアウト程度ないと強化学習が進まない。

# 対策

今回Egでは、ポリシーネットワークとして着手や探索数分布を学習する代わりに、リードネットワークとしてその手の価値～着手後のスコアを推定しています。

AlphaGoZeroやKataGoネットワークのポリシーヘッドでは、

$$L_{policy} = \pi^\top \log p$$

のように探索分布とポリシーを一致するよう学習するのに対して、

Egでは、自己対局中の教師局面で、探索した全ての手について、目数の推定値  $lead_m$  を保存し学習します。

$T$  を自己対局中に探索した手の集合、 $\alpha$ を探索しなかった手のペナルティとしたときに、

$$L_{mse} = \sum_{m \in T} (lead_m - p_m)^2$$

$$lead_{max} = \max_{m \in T} lead_m$$

$$L_{aux} = \sum_{m \notin T} ReLU(p_m - lead_{max} + \alpha)$$

$L_{mse}$  では教師局面で自己対局中に探索した手の目数差の誤差を最小化するよう学習し、

$L_{aux}$  では教師局面で自己対局中に探索しなかった手は、探索した最良の手より  $\alpha$  目以上悪いと仮定して未知の手の過大評価を防いでいます。

ポリシーネットワークの事前確率によらず同じ目数差になるので課題2に対応でき、ある局面ですべての全ての手の分布を知る必要が無いため課題3が発生しません。

## 実装と結果

探索プログラムとネットワークアーキテクチャはKataGoを流用しています。

RTX 4090 1台でランダム着手から強化学習し、17日でRay 8.0に勝ち越しました。

## UEC杯

KataGoの最新モデル(kata1-b28c512nbt-s7168446720-d4316919285)で生成した棋譜から強化学習を始めた大規模(b18c384nbt)モデルの強化学習しました。

大規模モデル: AMD EPYC 7R32 + NVIDIA A10G (AWS g5.12xlarge)