Reinforcement Learning On Connect4

Elias Sandin

Lund University

2022

Choosing Project

- EDAP01, Artificial Intelligence
- AlphaGo Zero paper
- Chess requires strong hardware

AlphaGo Zero

- Previous method used supervised learning
- Zero learns tabula rasa
- Policy network outputs action probability vector
- Value network estimates expected reward for state s

AlphaGo Zero

- MCTS executed for each position
- Selects moves that maximize $Q(s,a) + \frac{P(s,a)}{(1+N(s,a))}$
- Edges traversed in search update visit count, and action value
- Visit count $N(s,a)$ becomes target search probabilities
- data (s, π, z), minimize error of predicted value and winner, maximize similarity to π

Differences to previous AlphaGo

- No supervised learning from human data
- No rollouts
- Only boards as input

My version

- Started by pretraining policy and value network
- Test agains randomp player and filterplayer
- Then reinforcement learning with MCTS
- Leaf evaluation using rollouts + value network
- Rollout uses fast policy + randomness, relied more on rollouts in beginning
- Update data for all timesteps after termination

Data

- End nodes have value z
- Other nodes have value $\frac{1}{N(s_{t-1}, a_{t-1})} \sum_b N(s_t, b_t) \cdot Q(s_t, b_t)$
- Probabilities is visit percentage per edge

Some features

- Leaf node lambda depends on number of available moves at s' and depth of rollout
- Value updates for edges also depend on depth at obtained result
- Node satisfaction
  - if child C is terminal then parent node is satisfied
  - if there exists a satisfied child node with value 1 (win), parent node is also satisfied
  - if all children are satisfied parent is also satisfied
  - Parent value $= -max(childvalue)$
  - Used this to update data after game is finished

Reflections

- Can we tell if search probabilities are wrong?
  - When edges are satisfied "bad moves" can't be seen as bad if loss is inevitable
  - if not satisfied, maybe loss could be avoided
- When values are updated we're essentially updating probability of win in state s
- If new win probability is lower than previously thought, the difference could be seen as new observations