Metalevel RL for MCTS

Principles of Metalevel Control
by Nicholas Hay

Presented by Eric Langlois
Metalevel Monte Carlo Tree Search

Hex game image is public domain from wikipedia, the search tree is from the thesis, and the rest I created
Monte Carlo Tree Search

Image Source:
Pointed Trees

From the thesis; I repositioned the labels

\[ T_3 = \text{Tree}(N_3, T_{31}, T_{32}). \]

\[ T = \text{PointedTree}(C, N, T_1, T_2, T_3). \]

\[ C = \text{Context}(C', N_C, 1, T_{C1}). \]
Recursive Functions on Pointed Trees

$$f(T, f(\cdot), f(\cdot), f(\cdot), \ldots)$$

I created the formula display using copied bits of the tree image from the thesis.
Local Computations on Pointed Trees

Operations

- up(T)
- down(T, i)
- modify(T, N')
- insert(T, i, T')

f(Context)

f(Tree)

From the thesis
MCTs as a Metalevel MDP

- Simulation
- Expansion
- Simulation
- Backpropagation

Fixed Implementation with Pointed Tree Operations

MCTs figure is from another paper and Hex board is public domain image. I added the rest
Learning a Metalevel Agent

\[ \pi_\theta(T) = g_\theta(f_{\text{fixed}}(T)) \]

Tree Functions

- num_visits(T)
- all_done(T)
- average_rollout_value(T)
- average_estimate_value(T)
- p_over_n+1(T)
- minimax(T)

Context Functions

- depth(C)
- alpha(C) \textit{utility lower bound}
- beta(C) \textit{utility upper bound}
Policy Network $g_\theta$

$$\pi_\theta(T) = g_\theta(f_{\text{fixed}}(T))$$

Training Algorithm: TRPO with Generalized Advantage Estimation
John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel.
Highdimensional continuous control using generalized advantage estimation. ICLR, 2016
Experiments

Hex with a L x L board, maximum of n simulated actions per move

Average win rate against a UCT baseline

<table>
<thead>
<tr>
<th>L=3</th>
<th>n=10</th>
<th>n=20</th>
<th>n=50</th>
<th>n=100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.60</td>
<td>0.44</td>
<td>-0.59</td>
<td>-0.60</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.48</td>
<td>-0.42</td>
<td>-0.91</td>
</tr>
<tr>
<td></td>
<td>0.35</td>
<td>0.50</td>
<td>-0.63</td>
<td>-0.54</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>n=10</th>
<th>n=20</th>
<th>n=50</th>
<th>n=100</th>
</tr>
</thead>
<tbody>
<tr>
<td>L=3</td>
<td>0.58</td>
<td>0.47</td>
<td>0.47</td>
<td>0.33</td>
</tr>
<tr>
<td>L=5</td>
<td>0.70</td>
<td>0.60</td>
<td>0.51</td>
<td>-0.89</td>
</tr>
<tr>
<td>L=7</td>
<td>0.37</td>
<td>0.47</td>
<td>0.36</td>
<td>-0.45</td>
</tr>
</tbody>
</table>
Related Work

Metareasoning and Bounded Optimality

- Stuart J. Russell and Eric H. Wefald. Decision-theoretic control of search: General theory and an application to game-playing.

Monte-Carlo Tree Search