Modular multitask reinforcement learning with policy sketches

Jacob Andreas, Sergey Levine and Dan Klein
The learning problem

make planks
The learning problem

make planks

make sticks
Learning from sketches

- get wood
- use saw
- get wood
- use axe
The options framework
The options framework
The options framework

Diagram: A sequence of symbols indicating options and decisions, leading to a final outcome of +1.
The options framework

[Sutton et al. 99, Bacon & Precup 16]
Learning from intermediate rewards

[Kearns & Singh 02, Kulkarni et al. 16]
Learning from demonstrations

[Stolle & Precup 02, Fox & Krishnan et al. 16]
Learning from policy sketches

get wood \rightarrow use saw
Why sketches?

Easy to collect

<table>
<thead>
<tr>
<th>Task</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>make plank</td>
<td>get wood, use tools</td>
</tr>
<tr>
<td>make stick</td>
<td>get wood, use workbench</td>
</tr>
<tr>
<td>make cloth</td>
<td>get grass, use factory</td>
</tr>
<tr>
<td>make rope</td>
<td>get grass, use tools</td>
</tr>
<tr>
<td>make bridge</td>
<td>get iron, get wood</td>
</tr>
<tr>
<td>make bed*</td>
<td>get wood, use tools</td>
</tr>
<tr>
<td>make axe*</td>
<td>get wood, use workbench</td>
</tr>
<tr>
<td>make shears</td>
<td>get wood, use workbench</td>
</tr>
<tr>
<td>get gold</td>
<td>get iron, get wood</td>
</tr>
<tr>
<td>get gem</td>
<td>get wood, use workbench</td>
</tr>
</tbody>
</table>

Portable
Learning from policy sketches
Learning from policy sketches

- get wood
- use saw
- make planks
Learning from policy sketches

- get wood
- use axe
- make sticks
Learning from policy sketches

[Adapted from Branavan et al. 09, Oh et al. 17, Hermann et al. 17]

Example 1:
- **\( \pi_a \)**: "get wood then use saw"

Example 2:
- **\( \pi_b \)**: "get wood then use axe"
Learning from policy sketches

1. get wood
2. use saw
3. get wood
4. use axe
get wood

use saw

$\pi_1$

$\pi_2$

get wood

use axe

$\pi_1$

$\pi_3$
get wood

$\pi_1$
Policy representation

$\pi_1$  

get wood
Policy representation

$\pi_1$  

get wood
Policy representation
Policy representation
Policy representation
Policy representation

Action probabilities

\( \pi_1 \)

get wood
Policy search

\[
\sum_{\text{tasks}} \sum_{\text{steps}} (\nabla \log \pi(\rightarrow|\text{state})) (r_t - b)
\]
Policy search

\[ \sum \sum \left( \nabla \log \pi \left( \rightarrow \bigg| \begin{array}{c} \text{tasks} \\ \text{steps} \end{array} \right) \right) (r_t - b) \]
\[
\sum_{\text{tasks}} \sum_{\text{steps}} \left( \nabla \log \pi(\rightarrow) \right) (r_t - b)
\]
Policy search

\[ \sum_{\text{tasks}} \sum_{\text{steps}} \left( \nabla \log \pi \right) (r_t - b) \]

Reward

.40
Improving policy search
Improving policy search

\[ \sum \sum (\nabla \log \pi(\rightarrow|\text{state}))(r_t - b) \]
### Improving policy search

<table>
<thead>
<tr>
<th>Action</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use saw</td>
<td>Make planks</td>
</tr>
<tr>
<td>Use axe</td>
<td>Make planks</td>
</tr>
<tr>
<td>Get wood</td>
<td>Make planks</td>
</tr>
<tr>
<td>Get iron</td>
<td>Make planks</td>
</tr>
<tr>
<td>Use saw</td>
<td>Make nails</td>
</tr>
<tr>
<td>Use axe</td>
<td>Make nails</td>
</tr>
<tr>
<td>Get wood</td>
<td>Make nails</td>
</tr>
<tr>
<td>Get iron</td>
<td>Make nails</td>
</tr>
</tbody>
</table>
Improving policy search

\[ \sum_{tasks} \sum_{steps} \left( \nabla \log \text{SUBPOLICY} \right) (r_t - \text{TASK}) \]
Do sketches help?
The maze navigation task
The maze navigation task
The maze navigation task

Reward vs. x $10^6$ episodes

Sketches: modular
Unsupervised
Sketches: joint
The mini-craft task
The mini-craft task
The mini-craft task

Sketches: modular

Sketches: joint
Unsupervised

Reward

0 1 2 3 x 10^6 episodes
The cliff-walking task
The cliff-walking task

Sketches: modular

Sketches: joint

Unsupervised

log Reward

x $10^8$ timesteps
Zero-shot generalization

What if I see a sketch I’ve never seen before?

- get iron
- use axe
Zero-shot generalization

What if I see a sketch I’ve never seen before?
What if I see a sketch I’ve never seen before?
Fast adaptation

What if I don’t get a sketch at test time?

???
Fast adaptation

What if I don’t get a sketch at test time?

Unsupervised

Sketches

<table>
<thead>
<tr>
<th>Multitask</th>
<th>Adaptation</th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
<td>89</td>
</tr>
</tbody>
</table>
Fast adaptation

What if I don’t get a sketch at test time?

Unsupervised

Sketches

Multitask

Adaptation

47

89

42

76
Conclusions
A tiny bit of data goes a long way

The complete list of tasks, sketches, and symbols is given below. Tasks marked with an asterisk are held out for the generalization experiments described in Section 4.4, but included in the multitask training experiments in Sections 4.2 and 4.3.

<table>
<thead>
<tr>
<th>Goal Sketch</th>
<th>Maze environment</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>goal1 left left</td>
<td></td>
</tr>
<tr>
<td></td>
<td>goal2 left down</td>
<td></td>
</tr>
<tr>
<td></td>
<td>goal3 right down</td>
<td></td>
</tr>
<tr>
<td></td>
<td>goal4 up left</td>
<td></td>
</tr>
<tr>
<td></td>
<td>goal5 up right</td>
<td></td>
</tr>
<tr>
<td></td>
<td>goal6 up right up</td>
<td></td>
</tr>
<tr>
<td></td>
<td>goal7 down right up</td>
<td></td>
</tr>
<tr>
<td></td>
<td>goal8 left left down</td>
<td></td>
</tr>
<tr>
<td></td>
<td>goal9 right down down</td>
<td></td>
</tr>
<tr>
<td></td>
<td>goal10 left up right</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Goal Sketch</th>
<th>Crafting environment</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>make plank</td>
<td>get wood</td>
<td>use toolshed</td>
</tr>
<tr>
<td>make stick</td>
<td>get wood</td>
<td>use workbench</td>
</tr>
<tr>
<td>make cloth</td>
<td>get grass</td>
<td>use factory</td>
</tr>
<tr>
<td>make rope</td>
<td>get grass</td>
<td>use toolshed</td>
</tr>
<tr>
<td>make bridge</td>
<td>get iron</td>
<td>get wood</td>
</tr>
<tr>
<td>make bed*</td>
<td>get wood</td>
<td>use toolshed</td>
</tr>
<tr>
<td>make axe*</td>
<td>get wood</td>
<td>use workbench</td>
</tr>
<tr>
<td>make shears</td>
<td>get wood</td>
<td>use workbench</td>
</tr>
<tr>
<td>get gold</td>
<td>get iron</td>
<td>get wood</td>
</tr>
<tr>
<td>get gem</td>
<td>get wood</td>
<td>use workbench</td>
</tr>
</tbody>
</table>
Under review as a conference paper at ICLR 2017

AT

ASKS AND

SKETCHES

The complete list of tasks, sketches, and symbols is given below. Tasks marked with an asterisk (*) are held out for the generalization experiments described in Section 4.4, but included in the multitask training experiments in Sections 4.2 and 4.3.

Goal Sketch

Maze environment

goal1 left left

goal2 left down

goal3 right down

goal4 up left

goal5 up right

goal6 up right up

goal7 down right up

goal8 left left down

goal9 right down down

goal10 left up right

Crafting environment

make plank get wood
make stick get wood
make cloth get grass
make rope get grass
make bridge get iron
make bed* get wood
make axe* get wood
make shears get wood
get gold get wood
get gem get wood

A tiny bit of data goes a long way
Thank you!

https://github.com/jacobandreas/psketch