The Option-Critic Architecture

Pierre-Luc Bacon, Jean Harb, Doina Precup

Reasoning and Learning Lab
McGill University, Montreal, Canada

AAAI 2017
Intelligence:
the ability to generalize and adapt efficiently to new and uncertain situations

- Having good representations is key

“[…] solving a problem simply means representing it so as to make the solution transparent.” — Simon, 1969
Equipped with a good state representation, RL has led to impressive results:

- Tesauro’s TD Gammon (1995),
- Watson’s Daily-Double Wagering in Jeopardy! (2013),
- Human-level video game play in the Atari games (2013),
- AlphaGo (2016)...

The ability to abstract knowledge temporally over many different time scales is still missing.
Temporal abstraction

Higher level steps
Choosing the type of coffee maker, type of coffee beans

Medium level steps
Grind the beans, measure the right quantity of water, boil the water

Lower level steps
Wrist and arm movements while adding coffee to the filter, ...
Temporal abstraction in AI

A cornerstone of AI planning since the 1970’s:

It has been shown to:
- Generate shorter plans
- Reduce the complexity of choosing actions
- Provide robustness against model misspecification
- Improve exploration by taking shortcuts in the environment
Temporal abstraction in RL

*Options* (Sutton, Singh, Precup 2000) can represent courses of action at variable time scales:

**High level**

**Low level**

**Trajectory, time**
Options framework

An option $\omega$ is a triple:

1. initiation set: $\mathcal{I}_\omega$
2. internal policy: $\pi_\omega$
3. termination condition: $\beta_\omega$

Example

Robot navigation: if there is no obstacle in front ($\mathcal{I}_\omega$), go forward ($\pi_\omega$) until you get too close to another object ($\beta_\omega$)

We can derive a policy over options $\pi_\Omega$ that maximizes the expected discounted sum of rewards:

$$
\mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \bigg| s_0, \omega_0 \right]
$$
Contribution of this work

The problem of **constructing/discovering** good options has been a challenge for more than 15 years.

**Option-critic** is a scalable solution to this problem:

- Online, continual and model-free (but models can be used if desired)
- Requires no a priori domain knowledge, decomposition, or human intervention
- Learns in a single task, at least as fast as other methods which do not use temporal abstraction
- Applies to general continuous state and action spaces
The policy (actor) is decoupled from its value function.
The critic provides feedback to improve the actor.
Learning is fully online.
• Parameterize internal policies and termination conditions
• Policy over options is computed by a separate process
Main result: Gradient updates

• The gradient wrt. the internal policy parameters \( \theta \) is given by:

\[
\mathbb{E} \left[ \frac{\partial \log \pi_{\omega,\theta}(a|s)}{\partial \theta} Q_U(s, \omega, a) \right]
\]

This has the usual interpretation: take better primitives more often inside the option

• The gradient wrt. the termination parameters \( \nu \) is given by:

\[
\mathbb{E} \left[ -\frac{\partial \beta_{\omega,\nu}(s')}{\partial \nu} A_{\pi\Omega}(s', \omega) \right]
\]

where \( A_{\pi\Omega} = Q_{\pi\Omega} - V_{\pi\Omega} \) is the advantage function This means that we want to lengthen options that have a large advantage
Results: Options transfer

Hallways

Walls

Initial goal

Random goal after 1000 episodes
Results: Options transfer

- Learning in the first task no slower than using primitives
- Learning once the goal is moved faster with the options
Results: Learned options are intuitive

Probability of terminating in a particular state, for each option:

- Terminations are more likely near hallways (although there are no pseudo-rewards provided)
Results: Nonlinear function approximation

Same architecture as DQN (Mnih & al., 2013) for the 4 first layers but hybridized with options and the policy over them.
Performance matching or better than DQN

(a) Asterix (b) Ms. Pacman

(c) Seaquest (d) Zaxxon
Interpretable and specialized options in Seaquest

- Option 1: downward shooting sequences
- Option 2: upward shooting sequences

Action trajectory, time

White: option 1
Black: option 2

Transition from option 1 to 2

Option 1: downward shooting sequences
Option 2: upward shooting sequences
Conclusion

Our results seem to be the first to be:

- fully end-to-end
- within a single task
- at speed comparable or better than using just primitive methods

Using ideas from policy gradient methods, option-critic

- provides continual option construction
- can be used with nonlinear function approximators
- can incorporate regularizers or pseudo-rewards easily
Future work

- Learn initiation sets:
  - Would require a new notion of stochastic initiation functions
- More empirical results!

Try our code:
https://github.com/jeanharb/option_critic