Distral as variational information optimization

DJ Strouse

September 13, 2017

Abstract

[Teh et al, 2017] recently introduced an approach to transfer in a multi-task reinforcement learning setting. We show here that their approach is equivalent to regularizing agents with a variational bound on the mutual information between goals and actions given states.

1 A variational upper bound on $I(\text{goal};\text{action} \mid \text{state})$

We seek a variational (upper) bound on $I(G; A \mid S)$, where G is the goal, A is the action, and S is the state. We'll first develop an upper bound on $I(G; A \mid S = s)$, and then we'll average over $p(s)$ to get $I(G; A \mid S)$ afterwards. We begin by breaking up the mutual info into the difference of entropies:

$$\begin{aligned} I(G; A \mid S = s) &= H(A \mid S = s) - H(A \mid G, S = s) && (1) \\ &= \sum_{a,g} p(g \mid s)\, \pi_g(a \mid s) \log \pi_g(a \mid s) - \sum_a p(a \mid s) \log p(a \mid s)\,, && (2) \end{aligned}$$

where $\pi_g(a \mid s) \equiv p(a \mid s, g)$. Per the usual arguments, we assume that marginalizing over goals to get $p(a \mid s) = \sum_g p(g)\, \pi_g(a \mid s)$ is intractable, and so we approximate it with a varitional prior $\pi_0(a \mid s)$. Since $\mathrm{KL}[p(a \mid s) \mid \pi_0(a \mid s)] \geq 0$, we have $\sum_a p(a \mid s) \log p(a \mid s) \geq \sum_a p(a \mid s) \log \pi_0(a \mid s)$. Substituting into the above, we get the upper bound:

$$\begin{aligned} I(G; A \mid S = s) &\leq \sum_g p(g \mid s) \sum_a \pi_g(a \mid s) \log \frac{\pi_g(a \mid s)}{\pi_0(a \mid s)} && (3) \\ &= \sum_g p(g \mid s)\, D_{\mathrm{KL}}[\pi_g(a \mid s) \mid \pi_0(a \mid s)]\,. && (4) \end{aligned}$$

Now we average over state probabilities:

$$\begin{aligned} I(G; A \mid S) &\leq \sum_s p(s) \sum_g p(g \mid s)\, D_{\mathrm{KL}}[\pi_g(a \mid s) \mid \pi_0(a \mid s)] && (5) \\ &= \sum_g p(g) \sum_s p(s \mid g)\, D_{\mathrm{KL}}[\pi_g(a \mid s) \mid \pi_0(a \mid s)]\,. && (6) \end{aligned}$$

This suggests we can minimize (a variational upper bound on) $I(G; A \mid S)$ by sampling goals, sampling trajectories under that goal, and for each step, regularizing the agent with $D_{\mathrm{KL}}[\pi_g(a \mid s) \mid \pi_0(a \mid s)]$. This term can be optimized both with respect to the goal-specific policies $\pi_g(a \mid s)$, as well as the goal-independent "base policy" $\pi_0(a \mid s)$.

2 Distral

...and that's exactly what [Teh et al, 2017] do. Therefore, the only difference from our setup is that we explicitly parameterize the agent to produce a latent representation of the goal on the way to producing the policy, whereas they produce the policy outright. The advantage of their setup is fewer parameters; the advantage of ours is the ability to study the agent's goal representation directly. However, both are based on optimizing the same quantity, and thus I imagine their goal-specific and base policies will be more or less identical to our's.

References

[Teh et al, 2017] Yee Whye Teh, Victor Bapst, Wojciech Marian Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, Razvan Pascanu. *Distral: Robust Multitask Reinforcement Learning.* https://arxiv.org/abs/1707.04175