RL+Imitation

Paul Christiano Follow
Oct 15, 2016 · 4 min read

Reinforcement learning and imitation are two natural models for powerful AI systems. Both models have shortcomings:

- Unless we know what parts of behavior are “important,” an imitator needs to model *every* part of a behavior before it is guaranteed to get good results. Moreover, we can’t focus our model capacity or computational resources on the important aspects of behavior, and so we need to use a bigger model to get good performance.
- Reinforcement learning is extremely challenging. Exploration can take an exponentially long time, and only works when the problem is sufficiently nice (e.g. if the reward provides a trail of breadcrumbs leading to a good policy).

Another very natural AI problem is the intersection of imitation and RL: given an expert policy *and* a reward function, try to achieve a reward as high as the expert policy. The expert provides a general sense of what you should do, while the reward function allows you to focus on the most important aspects of the behavior.

In the same way that we can reason about AI control by taking as given a powerful RL system or powerful generative modeling, we could take as given a powerful solution to RL+imitation. I think that this is probably a better assumption to work with.

AI control with RL+Imitation

RL+imitation is a weaker assumption than either RL or imitation. In order to get any work out of it, we need to have access to both a reward function and demonstrations. So control schemes that work with RL+imitation are harder to design and more widely applicable—of course they can be applied if we have *either* an RL agent or to an imitation learner, but I also expect they are more likely to be applicable to some as-yet-unknown alternative capabilities.

Note that an imitation+RL system will *compete with the expert*, but we can’t assume that it will actually resemble the human behavior **or** that it will get high rewards. In order to argue that our AI will learn to do X we need to establish three claims:

- The expert policy does X.
- Doing X leads to a higher reward.
- The AI can learn to do X. (And doing X gives more reward than other uses of its model capacity.)

If we want to argue that our AI *won't* do X, then we need to establish all three of the opposite properties: our AI can learn to avoid X, the expert avoids X, and doing X leads to a low reward.

Examples

Human children often solve imitation+RL problems, using imitation to figure out basically what to do, but using feedback in order to refine their behavior and understand what aspects are important.

AlphaGo solves an imitation+RL problem, exploiting both human demonstrations and extensive self-play.

Evolution solves a pure RL problem; it has no access to demonstrations and must figure things out the hard way.

Miscellany

I'm sure that imitation+RL has been studied formally and I apologize for not citing the sources. It's not something I've worked on from a technical perspective, and I'm not aware of prior work.

Note that imitation+RL is an *easier* problem than RL, and a better-defined problem than imitation. It's hard to say whether it is "easier" than imitation because that depends on how we measure success at imitation.

Note that imitation+RL is consistent with either a model-based or model-free approach to RL.

As with imitation, any scheme to get superhuman performance out of imitation+RL is going to require capability amplification or something similar.

Paths to AI

I think that imitation+RL is a likely path to building human-level AI. It is a path that tries to steal from the work of biological evolution and cultural accumulation, continuing the trajectory of human technological and social development.

Reinforcement learning without imitation is a different plausible path to powerful AI. I expect that pure RL involves higher computational demands and would imply somewhat later AI. I'd also guess that pure RL is slightly worse news from a control perspective, though I don't think that trying to push the field one way or the other is a useful exercise from a control perspective.

Discovery

At face value it may look like imitation+RL cannot discover anything new, since it can only perform behaviors that an expert can demonstrate. But we can explicitly regard exploration and discovery as its own problem, and learn to pursue these goals as effectively as humans do. This does require generalizing across many different acts of discovery, since we want to build a system that finds new things rather than imitating the discovery of an old thing, but such generalization seems essential for powerful AI at any rate.

(This is essentially what happens when we get new behaviors from policy amplification.)

I am particularly partial to this view of exploration/discovery, because it is basically necessary for my approach to AI control—even if I had access to an RL agent, I would try to teach the RL agent to explore in a way that humans approve of, rather than trying to e.g. find new ways to think as part of novelty-based exploration.

Conclusion

Going forward, I'll preferentially design AI control schemes using imitation+RL rather than imitation, episodic RL, or some other assumption. I think this opens up some interesting new questions, will help make theory better match reality, and will make control schemes more broadly applicable.