Reward engineering

Paul Christiano Follow

Dec 3, 2015 · 2 min read

This post gestures at a handful of research questions with a loose thematic connection.

The idea

Consider the following frameworks:

- Temporal difference learning: learn to predict the future by predicting tomorrow's prediction.
- Generative adversarial models: learn to sample from a distribution by fooling a distinguisher.
- Predictability minimization: learn to represent data efficiently by making each part of the representation unpredictable given the others.

These algorithms replace a hard-to-optimize objective with a nicer proxy. These proxies are themselves defined by machine learning systems rather than being specified explicitly. I think this is a really nice paradigm, and my guess is that it will become more important if large-scale supervised and reinforcement learning continues to be a dominant methodology.

Following Daniel Dewey, I'll call this flavor of research "reward engineering." In terms of tools and techniques I don't know if this is a really a distinct category of research; but I do think that it might be a useful heuristic about where to look for problems relevant to AI control.

Relevance to AI control

Though reward engineering seems very broadly useful in AI, I expect it to be especially important for AI control:

- A key goal of AI control is using AI systems to optimize objectives which are defined implicitly or based on expensive human feedback. We will probably need to use complex proxies for this feedback if we want to apply reinforcement learning.

- Reward engineering seems relatively robust to changes in AI techniques. Uncertainty about future techniques is often a major obstacle to doing meaningful work on AI control in advance (even if only a little bit in advance).

Applications

I see a few especially interesting opportunities for reward engineering for AI control:

- Making efficient use of human feedback. Here we have direct access to the objective we really care about, and it is just too expensive to frequently evaluate. (*Simple proposal*: train a learner to predict human judgments, then use those predicted judgments in place of real feedback.)
- Combining the benefits of imitation and approval-direction. I suspect it is possible to avoid perverse instantiation concerns while also providing a flexible training signal. (*Simple proposal*: use the adversarial generative models framework, and have the operator accomplish the desired task in a way optimized to fool the distinguisher.)
- Increasing robustness. If our ML systems are sufficiently sophisticated to foresee possible problems, then we might be able to leverage those predictions to avoid the problems altogether. (*Simple proposal*: train a generative model to produce data from the test distribution, with an extra reward for samples that "trip up" the current model.)

In each case I've made a preliminary simple proposal, but I think it is quite possible that a clever trick could make the problem look radically more tractable. A search for clever tricks is likely to come up empty, but hits could be very valuable (and would be good candidates for things to experiment with).

Beyond these semi-specific applications, I have a more general intuition that thinking about this aspect of the AI control problem may turn up interesting further directions.