Turning reflection up to 11

Paul Christiano Follow
Feb 2, 2016 · 4 min read

Suppose that Hugh wants to build a super-Hugh-level question-answerer.

Hugh could train a system to imitate Hugh's answers. But that approach will at best be Hugh-level.

What Hugh really wants is a system that predicts how Hugh *would have answered*, if Hugh had thought about the question at great length. And by "great length," we mean "lengths so great that Hugh doesn't actually have the time to do it."

This would be nice, but there is no obvious way to get the training data. So what should Hugh do?

(Hugh could rate the system's answers, and train it to maximize the ratings. But Hugh really wants to maximize the ratings he *would have assigned* if he thought about the question at great length. In fact, that's the case where this technique really matters.)

Extrapolation

Here is a simple procedure. We'll have Hugh answer different questions in different amounts of time. For concreteness, let's define 8 levels of deliberation:

1. 10 seconds [Hugh gathers 1 million data points at this level of deliberation, spending one year full time on the problem.]
2. 1 minute [150,000 examples. Another year.]
3. 10 minutes [15,000 examples]
4. 1 hour [2,500 examples]
5. 1 day [350 examples]
6. 1 week [52 examples]
7. 2 months [6 examples]
8. 1 year [1 example]

Each (Q, A) pair is tagged with its level of deliberation *n*. Rather than mapping Q → A, Hugh builds a system that maps (Q, *n*) → A.

To get a really great answer to question Q, Hugh queries the system on the pair (Q, 11). Hugh hopes that this will be a prediction of what he would answer, if he spent 500 years thinking about it.

Could this work?

I doubt it.

Certainly existing and foreseeable machine learning techniques wouldn't be able to handle this kind of extrapolation.

But from a theoretical perspective, I don't think it's the kind of question we *should* be able to answer.

First of all, the most natural value to assign to (Q, 11) is whatever value would have appeared along with it in the training set, if it had somehow appeared in the training set. This is certainly *not* the result of Hugh thinking for 600 years! Indeed, it's not even clear what hypothetical we *want* Arthur to consider when extrapolating to 11. And we have no way at all to communicate which of the possible generalizations it should use.

Certainly this approach cannot tolerate adversarial errors: an error which only affects levels ≥ 10 will never be detected or corrected by training.

In order to get good results when extrapolating to 11, we will have to lean *extremely* hard on our regularization/prior, since we have literally no data to constrain the model in that regime. Even getting good results for category 6 would require truly heroic feats, which are far beyond existing techniques. Extrapolating to 11 would seem to require qualitatively new ideas and techniques.

For simple functions we might imagine extrapolating in this way by constructing a very good prior (e.g. assuming the model is linear, or finding a simple model that fits the data perfectly). But for messy functions that are hard to learn and tied up with the physical world, it doesn't look like this approach will scale.

Symbolic reasoning

We might be able to symbolically define "what Hugh would say if he thought about the question for 600 years," and ask a system to answer that question. We could provide the same labelled examples, but now they are serving as logical information rather than as the task definition.

I think this is a central example of what we want to get out of symbolic reasoning, from an AI control perspective. I keep this example in mind whenever I talk about logical uncertainty—if a technique could deliver "good" judgments about this question, that would make me more optimistic about AI control. And if a technique can't help us with questions like this one, then I don't expect it will be helpful.

Overall I'm weakly pessimistic about addressing this problem using symbolic reasoning. But at least it gives us the ability to specify which generalization we want. And it also provides a bunch of additional structure that may take some weight off our prior. It might work out, and that would be great.

(The existence of labelled training data in small cases may make the symbolic reasoning problem seem easier. But my goal is to do as well as any available algorithm. And in particular that means that we must take advantage of whatever inductive regularities would be used for the simple supervised learning problem. It seems very hard to bridge the gap between supervised learning and the kind of logical structure that we need to use to constrain our beliefs about more complex sentences. That's the part that seems really hard to me.)

Interpolation (or: why this is useful anyway)

If we actually want to train a model to predict thoughtful human responses, it is probably helpful to collect many less-thoughtful responses as well. Less thoughtful responses are radically cheaper, but still capture a ton of information about the problem.

We could use the procedure from the last section, but limit ourselves to interpolation rather than extrapolation. (Incidentally, I think this is how AlphaGo aggregates data across experts of different strength, it's at least how they did it here.) Of course if we are good at semi-supervised learning, we can just hand Arthur the quick responses as unlabelled data and let it do its thing.

These are among the many techniques that seem worth exploring to reduce the burden of human oversight. This is basically the usual transfer learning problem (though I think that researchers interested in AI control can get extra mileage by thinking about the specific problem instances that are most relevant to AI control). Transfer learning already raises many hard questions.

Maybe if we answered those questions we would have some insight into whether and when extrapolation can work. But for now, I remain skeptical.