AlphaGo Zero and capability amplification

Paul Christiano Follow
Oct 19, 2017 · 3 min read

AlphaGo Zero is an impressive demonstration of AI capabilities. It also happens to be a nice proof-of-concept of a promising alignment strategy.

How AlphaGo Zero works

AlphaGo Zero learns two functions (which take as input the current board):

- A prior over moves **p** is trained to predict what AlphaGo will eventually decide to do.
- A value function **v** is trained to predict which player will win (if AlphaGo plays both sides)

Both are trained with supervised learning. Once we have these two functions, AlphaGo actually picks it moves by using 1600 steps of Monte Carlo tree search (MCTS), using **p** and **v** to guide the search. It trains **p** to bypass this expensive search process and directly pick good moves. As **p** improves, the expensive search becomes more powerful, and **p** chases this moving target.

Iterated capability amplification

In the simplest form of iterated capability amplification, we train one function:

- A "weak" policy **A**, which is trained to predict what the agent will eventually decide to do in a given situation.

Just like AlphaGo doesn't use the prior **p** directly to pick moves, we don't use the weak policy **A** directly to pick actions. Instead, we use a capability amplification scheme: we call **A** many times in order to produce more intelligent judgments. We train **A** to bypass this expensive amplification process and directly make intelligent decisions. As **A** improves, the amplified policy becomes more powerful, and **A** chases this moving target.

In the case of AlphaGo Zero, **A** is the prior over moves, and the amplification scheme is MCTS. (More precisely: **A** is the pair (**p**, **v**), and the amplification scheme is MCTS + using a rollout to see who wins.)

Outside of Go, **A** might be a question-answering system, which can be applied several times in order to first break a question down into pieces and then separately answer each component. Or it might be a policy that updates a cognitive workspace, which can be applied many times in order to "think longer" about an issue.

The significance

Reinforcement learners take a reward function and optimize it; unfortunately, it's not clear where to get a reward function that faithfully tracks what we care about. That's a key source of safety concerns.

By contrast, AlphaGo Zero takes a policy-improvement-operator (like MCTS) and converges towards a fixed point of that operator. If we can find a way to improve a policy *while preserving its alignment*, then we can apply the same algorithm in order to get very powerful but aligned strategies.

Using MCTS to achieve a simple goal in the real world wouldn't preserve alignment, so it doesn't fit the bill. But "think longer" might. As long as we start with a policy that is close enough to being aligned —a policy that "wants" to be aligned, in some sense—allowing it to think longer may make it both smarter *and* more aligned.

I think designing alignment-preserving policy amplification is a tractable problem today, which can be studied either in the context of existing ML or human coordination. So I think it's an exciting direction in AI alignment. A candidate solution could be incorporated directly into the AlphaGo Zero architecture, so we can already get empirical feedback on what works. If by good fortune powerful AI systems look like AlphaGo Zero, then that might get us much of the way to an aligned AI.