Imitation and justification

Paul Christiano Follow
Nov 6, 2015 · 4 min read

Suppose that I am training an AI system to play Go. One approach is to have the AI observe human moves and learn to predict those moves. The AI can then pick moves by sampling from its predicted distribution over "what a human would do."

But an AI may be able to learn more quickly by reproducing justifications along with the moves themselves. That is, we can use the modified training procedure:

- Each time a human makes a move, they provide a justification for that move. For example, the human may point out which groups are dead, or that a particular piece is a ladder breaker.
- The AI is trained to reproduce moves+justifications. (The justifications might have to be adjusted in order to be learnable.)
- In order to make a move, the AI produces a move+justification and then throws away the justification.

Naturally, this approach can be generalized to arbitrary tasks, not just playing games.

Benefits

The justification helps teach the AI system how to think about the problem, and so can accelerate a learning process which would otherwise have to learn a lot of hidden structure on its own. For example, it may be quite difficult to learn about ladder breakers by watching a sequence of games in which no ladders are ever played, while it is much easier if the justification is pointed out. This is much more similar to how individual humans learn new tasks—commentary from teachers is an essential part of the process.

Providing justifications can also make the AI system significantly easier to understand and correct—e.g. if the AI plays poorly because it mistakenly believes a live group is dead, it may explain this fact as part of its justification.

Depending on how we define "imitation," this process may have an additional advantage. Suppose that we define imitation by introducing an evaluator who attempts to distinguish human behavior from the AI's behavior, and training the AI to fool the evaluator. Then hearing justifications is helpful for the evaluator as well as the imitator, and so increases the rate of learning through this additional channel. Beyond hastening learning, explanations may reduce the probability of blatantly inhuman behavior, by eliminating cases where the evaluator can't understand why the human is doing what they are doing. This allows the evaluator to reliably give a low score to complex actions that don't serve the kinds of goals the human would claim to pursue.

Gathering training data, and counterfactual oversight

This proposal is consistent with gathering training data continuously rather than during a separate training phase.

With small probability, each time the AI would act it instead outputs its move + justification (but doesn't take any action). The human then provides an example move + justification, and a distinguisher tries to tell the difference. Our AI is trained to make the distinguishing task as hard as possible.

Bootstrapping and argument

I previously proposed that AI systems gather training data through an argument between an action-proposer and a reviewer. The reviewer can search for the parts of the proposed action (or its justification) which are most likely to be problematic, and then the human can provide additional training data by arbitrating the dispute —that is, by judging whether the most-likely-to-be-problematic aspects really are problematic. This kind of active search for possible mistakes seems very helpful for coping with realistic limitations of our learning systems.

The same dynamic can be applied to imitation. Consider the following game, played between two learning systems Alice and Bob:

- Alice proposes an action.
- Alice justifies her proposal.
- Bob criticizes this justification. Alice and Bob interact, carrying out an informal argument.

The argument is never evaluated formally, and no one "wins." Instead, the goal of the two players is to imitate human performance. That is, a human may decide to play the role of one of Alice or Bob, while the other is still played by an AI. As before, the goal of Alice is to behave in the same way that a human would in her shoes, and the goal of Bob is to behave in the same way that a human would in his shoes.

An evaluator can observe the entire transcript before trying to predict which players (if any) were human or machine. As with justifications, this can make the task of the evaluator much easier.

Fooling the evaluator can subsume the goal of "winning the argument:" if the human always wins the argument (as a human would judge "winning") then the evaluator can learn to use "winning the argument (according to a typical human observer)" as a signal about which player is human. So if the human tries to win the argument, then so will the AI systems.

To obtain a move, we just stop the interaction as soon as Alice proposes a move—during normal operation, Bob never does anything. His role is only to help train the system.

Comparison to approval-maximization

This proposal seems to have most of the advantages of approval-directed behavior, while having a minimal risk of perverse instantiation.

The key challenge is that imitating human behavior may be more difficult than actually solving the problem at hand. The human who is modeling the behavior can try to help, but it's not clear whether/when that will be enough. Hopefully other techniques can further bridge the gap, or we can develop a better understanding of how the human model can reliably make themselves imitable.

I suspect that practical approval-directed systems will *not* have a serious difficulty with perverse instantiation (for the reasons given here). But it's still a problem to keep in mind, and I think that trying to address the key challenge with imitation is the most straightforward way to attack the problem of perverse instantiation.