Learning representations

Paul Christiano Follow
Dec 2, 2015 · 4 min read

Many AI systems form internal representations of their current environment or of particular data. Practical act-based agents will also need to form such representations, but would do so in a different way. This looks like a good challenge problem for act-based approaches, and I suspect it will be useful for evaluating the feasibility of the approach.

I'll discuss approval-directed agents in this post for concreteness but the discussion is generally applicable.

The problem

Consider a neural network learning to caption images. A typical approach to this problem involves building an internal representation of a scene and then using that representation to predict how well a sentence describes the scene (which itself involves updating the representation as we see words from a proposed description). The representation can be trained by gradient descent, adjusting encoding and decoding at the same time in order to optimize performance on the end-to-end task. It may also be pretrained e.g. by first finding a representation that works for a simpler task, but this is not always needed.

Ideally an approval-directed agent would use approval "all the way down," including for building these internal representations. This is especially important when the representations are used for persistent state, but it would be nice to use approval-direction even in the labeling setting (and it is useful as a simple test case even if we cared only about learning representations of persistent state).

This requires getting human feedback on this intermediate representation, and optimizing that feedback rather than the ultimate performance of the algorithm.

This presents two big problems:

- Getting feedback is very expensive. We would need technical progress to even try to use data-hungry machine learning

methods this approach. (See discussion here.)

- The human has to understand the learned representation and the rationale for the representation.

Why care?

We might be tempted to treat persistent state as a special case, using next-step approval to guide actions but allowing the state to be optimized on the basis of future approval. I think that it is better to try and learn state in an approval-directed way.

Our AI systems will need to perform other tasks involving similarly complex planning and design. For example, building a factory involves producing preliminary designs which are then analyzed in more detail, modified, and finally tested and implemented. If we cannot handle state, then it seems likely that we will have difficulty handling many of these other tasks.

Learning representations is an especially useful instance of the more general problem, because it is one that is very easy to study today.

On top of that, treating state as a special case can in principle lead to problems, and I would prefer try to flesh out a theoretically robust solution. Failing that I would prefer to thoroughly understand the obstructions that make a theoretically robust solution impossible.

Comparison to existing techniques

- Approval-directed representation-learning generalizes handcrafted representations. Using handcrafted features is not fashionable at the moment, but it's certainly not completely impractical. It's easy to imagine that there are productive intermediate approaches that involve collaborations between human feature engineers and automated searches. Is training on end-to-end performance essential to getting good performance? Maybe, but it's not clear.

- Many approaches to unsupervised learning fit within the approval-directed framework to varying extents, by designing representation scoring functions that proxy for operator approval better than end-to-end backpropagation (at least in the initial stages of training).

These comparisons suggest that the approval-directed approach isn't totally outlandish. Both of them attack the "human understands something about the representation" problem head on, but dodge the training data problem by using simple proxies for human approval. In some sense it would be nice to train more interesting criteria based on human feedback, but these simple proxies also seem compatible with the act-based approach.

Much modern supervised learning uses or is at least compatible with unsupervised pre-training, so in some sense the question is just: can we make the final end-to-end fine-tuning of a representation unnecessary or nearly superfluous? Normally pre-training mostly just points us in the right direction, but can we improve it far enough that it is doing much of the work? In practice the main reason people will care about this question is a lack of labelled data for the end-to-end fine-tuning (or maybe pure scientific interest), which are fine motivations.

In other domains, especially RNNs (whose internal state is not just about "representation" but can be intimately tied up with the encoding of a policy), the situation is trickier. There end-to-end training seems especially essential, and significant progress would be needed to even start living without it.

Upshot

I think that maintaining state is useful to have in mind as a challenging case for act-based agents, and in the long run if this approach is to succeed it will probably have to deal with the problem head-on.

Trying to reproduce results based on RNNs without using end-to-end training seems like a potentially interesting challenge problem (though it looks very hard and is definitely not at the top of my list).

Improving unsupervised feature learning, and exploring the intermediate space between handcrafted features and unsupervised learning, seems like an interesting domain. But it doesn't seem like an especially promising place for a researcher interested in AI control to focus, given that these problem are likely to receive lots of attention for other reasons.