Efficient feedback

Paul Christiano Follow
Nov 24, 2015 · 7 min read

In some machine learning domains, such as image classification, we can produce a bunch of labelled training data and use the same data to train many models. This paradigm is very efficient, but it's not always applicable. For example:

- Rather than imitating human decisions, we may want to train a system to make decisions that a human would *evaluate* favorably. If the space of possible outputs is large, feedback will effectively be specific to a particular algorithm that is being trained.
- If a learning system interacts with a complex environment, different algorithms will shape their environments differently and find themselves in different situations—requiring different training data.

These situations are quite natural, but are hard to address with the usual paradigm. This is a problem if we are interested in applying data-hungry algorithms to domains with these characteristics. In this case we may need to collect a lot of new data whenever we want to train a new model, or even multiple times during the training of a single model.

This is an especially important challenge for implementing counterfactual oversight; I think it's also an important barrier for implementing many practical AI control projects today.

Outline

In section 0 I'll introduce a running example and explain the problem in slightly more detail.

In section 1 I'll discuss some plausible approaches to this problem.

In section 2 I'll discuss the relevance to AI control, and describe some possible domains for studying the problem.

In section 3 I'll describe how I think that the proposed research differs from existing research in similar directions.

0. Example

Suppose that I am training a question-answering system, which I hope will map natural-language questions to acceptable natural-language responses.

The most common approach is for humans to provide a bunch of answers in a labelled database of (Q, A) pairs. This approach is not completely satisfactory:

- This approach seems to throw out almost all of the actual structure of the task, by giving all answers unlike the human answer a payoff of 0. We can restore some of this structure, for example by introducing a similarity metric and rewarding similar answers, or by providing several candidate answers for each question. But these measures are very incomplete, and they require considerable domain-specific tweaking.

- This training forces our system to answer questions in a human-like way, rather than answering them in whatever way it can. For example, this system won't learn to: use formulaic and mechanical language, even if doing so is its best way to give clear and understandable responses; outperform the human reviewers in domains it finds easy; express ignorance about common-sense facts rather than making a guess; give an OK but obviously incomplete answer when it can't produce a good one…

- If we want to use this system to generate entire dialogs, then it may end up answering questions from a distribution that is very far from the training data. For example, users may ask clarifying questions that are very uncommon in normal human discussions. Or users may respond to a system's errors by providing very explicit and detailed information about the particular kinds of errors it is making. We would like to train systems to behave appropriately in the real environment it will encounter.

Alternatively, we could train our system by providing *feedback*: we ask a question Q, it provides an answer A, and we score that answer. This allows us to score the kinds of answers it actually provides, and to adjust the distribution of questions based on the behavior of the algorithm (e.g. we could train on questions from actual interactions between our system and humans).

One salient difficulty with this approach is that it requires a lot of data specific to the algorithm we are training—we can’t build a large database of (Q, A) pairs and then use it for each new algorithm we want to train. Moreover, as our algorithm changes, it starts producing different answers. In order for the training process to continue, we need to continuously provide new feedback.

Acquiring all of this data seems expensive; that expense is a constraint on the kinds of techniques that we can practically pursue, and I think it’s a particularly hard constraint for AI control.

1. Techniques

This section lists some approaches to the problem of efficiently using expensive human feedback. Two of these approaches are standard topics of ML research. My view is that (1) additional research in these areas will have meaningful benefits for AI control, (2) researchers interested in AI control would pursue a distinctive angle on these questions [see section 3], and (3) confronting these issues in the context of intended AI control applications [see section 2] will be especially useful.

To the extent that existing techniques are good enough for the applications in section 2, it would be better to directly apply existing techniques. This would be good news for research on AI control—unfortunately, I don’t think that it is yet the case, and so some additional research focused on these issues is probably needed.

Learning to give feedback

In the question-answering setting, the user supplies a rating for each proposed answer.

One way to reduce human involvement is to train an evaluator to predict these ratings. That evaluator can be used to train the underlying question-answering system, with these two training processes proceeding in parallel.

In some sense this is a very straightforward approach, but I suspect that actually making it work well would be both challenging and informative.

It's not clear if it would be best to train the system to produce absolute scores, or to judge the relative merit of several proposed answers (which could also be used to construct a training signal). The advantage of comparisons is that we may want e.g. our rankings to become more strict as the system improves. The same model of comparisons can be used even as the evaluated algorithm becomes more sophisticated, while scores would need to be continuously adjusted.

Active learning

Rather than eliciting training data in every case or in a random subset of cases, we would like to focus our attention on the cases that are most likely to be informative, utilizing human input as effectively as possible. This is a standard research problem in machine learning.

Semi-supervised learning

In practice, our algorithms will have access to a lot of labelled and unlabelled data, in addition to information from human feedback. Efficient algorithms will have to combine these data sources. This is also a standard research problem.

2. Applications / motivation

Why would solving this problem be useful for AI control? I have two motivations:

1. Different training approaches involve different amounts of human interaction. I think that more interaction tends to be preferable from a control perspective, and I see a number of particular approaches to the control problem that are very interaction-heavy. So improved techniques for efficient human involvement have direct relevance to control, by improving the viability of these techniques relative to alternatives that involve less interaction.

2. The expense of human interaction already seems to be a serious obstacle to concrete research on many aspects of the AI control problem. In addition to being evidence that mechanism [1] is real, this gives a very practical motivation for mitigating improving interaction: it would help us study scalable mechanisms for AI control today.

For both purposes, it seems especially useful to study these issues in the context of scalable AI control mechanisms. This section describes two such domains.

Apprenticeship learning

The first two problems discussed in this post require feedback during training; making either of them work would likely require some of the techniques described here.

More generally, imitating human behavior in complex domains may require querying the humans on-policy. This issue is discussed and examined empirically in Ross and Bagnell 2010.

Explanation

From the AI control perspective, another natural problem is *explanation*—training systems to produce explanations of their own behavior. The relevance to AI control is discussed in this post.

Many researchers are interested in extracting explicable decisions from learned models, but I am not aware of any work that uses supervised learning to drive that process. The expense of human feedback seems to be a primary difficulty. Such feedback looks necessary, given that:

- the space of possible explanations is extremely large,
- the particular kind of explanation needed may depend on the algorithm which is trying to explain itself, and
- the explanations generated by humans and machines may be very different.

Actually training models to produce explanations would no doubt reveal many additional problems, but the training data issue makes it hard to even begin work except in toy cases.

3. Comparison to existing research on these topics

I would expect research in this direction to be contiguous with traditional ML research on semi-supervised and active learning. Nevertheless, there are some possible differences in focus. As opposed to traditional research in these areas, research targeting AI control might:

- Pick application domains of more direct relevance to AI control, such as explanation and apprenticeship learning.
- Explore the dynamic in “learning to give feedback,” and generally cope with difficulties distinctive to feedback as opposed to labels.
- Cope with a limited quantity of feedback / “on-policy” labels, while having ample access to traditional labelled data (which is usually the scarce resource for active or semi-supervised learning). In some sense this is more like off-policy reinforcement learning than like semi-supervised or active learning in a classification setting.
- Experiment with different kinds of feedback to find whatever works most efficiently, rather than taking the partial labelling or query model as part of the problem statement.
- Prefer “scalable” approaches, which are not too sensitive to details of the underlying algorithms, and which will become increasingly efficient as algorithms improve.

Conclusion

Some scalable approaches to AI control involve extensive user feedback. The cost of such feedback may be a long-term obstacle to their applicability, and at the moment it certainly seems to be an obstacle to experimenting with those approaches.

I think there are many promising research directions to make this feedback more efficient. These look like a good way to push on AI control, especially if pursued in domains that are directly relevant for control and with a focus guided by that application. I’m optimistic that this is another possible point of alignment between existing AI research and research on AI control.