Cyc and Machine Learning

Samuel Teeter
Member Of The Technical Staff
Cycorp, Inc.

November 26, 2019

1 Cyc vs ML

When modern computer scientists think of AI, they tend to think of powerful neural networks inferring patterns from millions of rows of data. These statistical learning methods, known collectively as machine learning or ML, have enjoyed enormous popularity in recent years due to the availability of large sets of training data and high performance computing environments. While machine reasoning systems such as Cyc are often grouped together with machine learning systems under the broad heading of "Artificial Intelligence," they are actually very different methods with different advantages and drawbacks depending on the type of problem they are trying to solve. This article will explain the key differences between Cyc and popular machine learning methods and outline the advantages and disadvantages of each approach to AI.

1.1 Small Data vs Big Data

One term commonly used in connection with machine learning is "big data." Machine learning methods such as neural networks are capable of inferring incredibly complex patterns within observed data points; but in order to guarantee a high probability of success for their models, they need a large number of data points to train on. "Big data" refers to the use of massive datasets, such as online product reviews, page hits, or social media activity, to train statistical models. This approach is useful for companies that need to make general, probabilistic judgments based on large datasets. Google's page ranking algorithm, Netflix's content recommendations, and Amazon's suggested products are all examples of this.

Other problems, however, require intelligent behavior without large-scale data analysis. A hospital, for example, might want to know how many staff members it needs to have onsite to meet the projected needs of its patients; or a supply chain manager might want to know how large of an order they can fill based on factors affecting the productivity of factories. These problems differ from the machine learning applications in two key ways:

1. The rules governing the problem are already understood. We don't need novel statistical models to discover how much care our patients require or how our supply chain operates; rather, we want to apply our knowledge of these well-studied problems at high scale and efficiency.

2. High accuracy is preferable to high applicability. A machine learning method can usually make a prediction for any input; but if that input is not similar to its training data, the prediction may have a very low probability of being correct. The questions posed above require strong answers that can be traced back to salient lines of reasoning.

1.2 Why vs What

Machine learning methods use mathematics to model data; this means that they can make predictions about a wide variety of datasets, but they cannot justify their predictions except through probabilistic reasoning. When a movie streaming website recommends a title to a viewer, they are effectively telling the viewer, "Your data sample is near to other viewers who liked this movie in the space of viewing history profiles."

Statistical reasoning makes ML methods very efficient, but it limits the complexity of our interactions with them. A viewer who enjoys artistic films but doesn't want to view titles by Woody Allen, for example, can't explain this to a statistical model. A Cyc-based recommendation system, however, could trace its recommendations back to specific observations about movies and viewers, allowing the user to make specific exceptions.

1.3 Knowledge vs Information

Machine learning thrives on data that is relatively unstructured and easy to represent numerically. Image processing is a prime example of this. By representing images and their labels as huge matrices of numbers, neural networks can learn to identify visual patterns that correspond to objects.

Machine learning tends to be weaker, however, with data that is structured and non-numerical. A good example of this is natural language processing. By representing words as vectors in a numerical feature space, machine learners can discover statistical relationships within language–but they have no representation whatsoever of what the language actually means. This is why chatbots tend to rely on commonplace figures of speech that often don't make any sense. A symbolic AI such as Cyc, on the other hand, contains logical concepts that actually represent the meaning of language, and can be used to compute a meaningful response.

1.4 Old vs ... Equally Old?

One common misconception about symbolic AI is that it is an "antiquated" approach to AI with little practical application. However, machine learning

methods are in fact equally old. Research on neural networks began in the mid-1960s, while knowledge-based systems became a popular research subject in the early 1970s. Since then, the speed of computing machinery has increased exponentially while the cost has dropped, and large-scale sets of training data have become widely available. This combination of factors has helped machine learning to flourish in recent years and led to a renewed interest in artificial intelligence.

These advances have also benefitted Cyc. With modern high-performance hardware, Cyc can perform inferences on a knowledge base millions of times larger than the one it started with. Just as machine learning methods take advantage of widely available unstructured datasets, knowledge-based systems can now exploit structured datasets such as Wikidata and WordNet. Cyc's ability to communicate with a wide variety of data sources enables it to quickly absorb knowledge about any topic, allowing it to solve a wider range of problems than ever before.

2 AI and Air Travel: Broad Categories

In summary, critics of knowledge-based reasoning often portray it as antiquated and simplistic, like a bi-plane compared to a modern jet. In fact, a better analogy would be the difference between a jet and a helicopter. Both are sophisticated tools that have evolved over many years of research; both are well-adapted for solving different kinds of problems; and both happen to be lumped together under a fairly broad category of engineering. Ultimately, the question of which approach is "better" depends on the nature of the problem at hand.

A more interesting question is what the two approaches can accomplish in tandem. Machine learners with a structured representation of knowedge can deliver better answers to semantic queries, as in the case of Google's Knowledge Graph. Cyc, by incorporating machine learning methods into its internal processes, can solve logical problems with greater efficiency and incorporate machine-learned facts into its knowledge base. At Cycorp, we often refer to this as the left-brain/right-brain concept. Machine reasoning, like the left brain, provides structure and logical reasoning; machine learning, like the right brain, provides flexibility and associativity. Together, they form a more complete and useful basis for understanding the world and solving problems.