Artificial Intelligence: A Guide for Thinking Humans（Chapter 9&10）

Melanie Mitchell is a professor of computer science at Portland State University. She has worked at the Santa Fe Institute and Los Alamos National Laboratory. In 2019, she published **<Artificial Intelligence: A guide for Thinking Humans>.**

This book is published by Farrar, Straus and Giroux(FSG), an American book publishing company. FSG is known for publishing literary books and its authors have won numerous awards, including Pulitzer Prizes and Nobel Prizes. FSG is undoubtedly a competitive publisher.

I personally very like these two chapters of Melanie Mitchell's book. Games and reinforcement learning are both interesting for me. Not only does Mitchell explain the main approaches clearly, her account is readable and engaging. There are theory parts in these two chapters but they are understandable.

The writing is clear and the texts are well structured. In Chapter 9&10 of this book, the author mainly introduced the application of reinforcement learning in games and Go. The author took AlphaGo as an example to show the potential of reinforcement learning in certain domains. There are also the doubts and author's thinkings about the future of reinforcement learning.

What is most important for me is that the author pointed out that current reinforcement learning models cannot do 'transfer learning' which refers to the ability of a program to transfer what it has learned about one task to help it perform a different, related task. This actually implies that current reinforcement learning models like AlphaGo are not really as same intelligent as humans. This is important because if current reinforcement learning models cannot do 'transfer learning', they cannot apply to real-world problems.

I agree with Mitchell when Mitchell wrote "how to think logically, reason abstractly, and plan strategically" is the real intelligence humans chase for. Because these things are the real reason that we think Go/Chess can improve our intelligence, not a simple win or loss.

But are reinforcement learning models really incapable of 'transfer learning'? Curious about that, I found one another article.

Transfer in Reinforcement Learning via Shared Features

<Transfer in Reinforcement Learning via Shared Features> has 3 authors: George Konidaris is a professor of Computer Science and Artificial Intelligence Laboratory in Massachusetts Institute of Technology; Ilya Scheidwasser is a professor of Department of Mathematics in Northeastern University; Andrew G. Barto is a professor of Department of Computer Science in University of Massachusetts Amherst.

This article is published on Journal of Machine Learning Research in 2012. The Journal of Machine Learning Research(JMLR) is a peer-reviewed open access scientific journal covering machine learning. Its h-index is 221 and its impact score is 5.41. This is a highly competitive journal.

This article is very clear and well structured: the authors firstly introduce reinforcement learning and notion of transfer; then introduce the framework for transfer along with some experiments; finally discuss the implications and limitations of their work.

To be frank, I do not like this article because in this article there are too many theoretical, obscure passages for me. Although it's a good thing for me to learn some new things, I found it a little hard to finish reading the whole article. However I do have to admit that the idea of this article is really genius and interesting.

The most important idea of this article is that related tasks share some common features, and that transfer can be achieved via those shared features in reinforcement learning. Based on this idea, the 3 authors presented a framework for transfer in reinforcement learning. The framework can capture the features of different but related tasks, and it provides some insight into when transfer can be applied to a problem and when it cannot. This is very important because this provides a practical possibility to create agents that can improve their own problem-solving capabilities through experience of multiple similar problems. In other words, the ability of 'transfer learning' mentioned above is not totally impossible if we work under such framework.

I personally agree with this idea. Imagine a simple question: Why many tennis players find ping-pong very easy to learn? Because some features of tennis and table-tennis are all the same(or at lease we could say, similar). Meanwhile, the results of experiences in this article also shows the possibility of 'transfer learning' in reinforcement learning domain.

We should make reinforcement learning models capable of transfer learning!

So, are reinforcement learning models really incapable of doing ‘transfer learning’? This is the most biggest difference between the two texts above. Mitchell’s answer is NO while Konidaris(and his coworkers) presented a framework which allows reinforcement learning models to transfer in some degrees. My point of view is that reinforcement learning does not guarantee the ability of transfer. We should make reinforcement learning models capable of transfer learning!

As everyone know, reinforcement learning’s inspiration comes from the theory of behaviorism in psychology: the agent will learn how to take actions in an environment in order to maximize the notion of cumulative reward. Reinforcement learning is just similar to humans’ learning process.

But ‘transfer’ is actually an another concept. Let us get back to the tennis and table-tennis example. We could transfer from tennis to table-tennis because these two sports have some common things. When we humans learn something and do ‘transfer’, we will also try to find the similar features and think about what I can use in our experience to behave well for new things. In this way, we could find that the definition of current reinforcement learning does not contain that part of ‘finding similar features’. That’s why I think reinforcement learning does not guarantee ‘transfer’. They are actually 2 different things.

In Chapter 5 <Transfer in Reinforcement Learning: A Framework and a Survey> of **<Reinforcement Learning>**, Alessandro Lazaric wrote that ‘Unlike supervised learning, reinforcement learning problems are characterized by a large number of elements such as the dynamics and the re- ward function, and many different transfer settings can be defined depending on the differences and similarities between the tasks.’ Reinforcement learning has a great potential to be combined with certain frameworks so that we can allow our RL models to do transfer learning.

And we should do more further work on this domain! Because the reason why Konidaris(and his coworkers) did such a fancy work is just like what Mitchell quote from Demis Hassabis(founder of DeepMind) in her book: ‘They finally can apply to real-world problems and have a huge impact on things like healthcare and science’. Once the reinforcement learning models can obtain the ability of ‘transfer’ in general, we could say that the time of AI really comes.

I didn’t find a relation between the two texts above on Google Scholar. **<Artificial Intelligence: A guide for Thinking Humans>** is published after **<Transfer in Reinforcement Learning via Shared Features>** but the former does not refer to the latter. Moreover, there are neither articles that quote both these two texts above nor articles that these two texts both quote.

I think this is because: athough they do have some similar and opposing ideas, **<Artificial Intelligence: A guide for Thinking Humans>** belongs to popular science book while **<Transfer in Reinforcement Learning via Shared Features>** is a scientific paper published in journal.