Reading Assignment from Melanie Mitchell's "Artificial Intelligence: A Guide for Thinking Humans"

Assignment by Carrie Corcoran

Chapter 6: A Closer Look at Machines that Learn

1. TRUE/FALSE - The learning-from-data approach of deep neural networks has generally proved to be more successful than the "good old-fashioned AI" strategy, in which human programmers construct explicit rules for intelligent behavior. However, contrary to what some media have reported, the learning process of ConvNets is not very humanlike.
True.

2. Why does your professor like the previous question?
This question introduces the topic of the limitations of ConvNets.

3. TRUE/FALSE - As we've seen, the most successful ConvNets learn via a supervised-learning procedure: they gradually change their weights as they process the examples in the training set again and again, over many epochs (that is, many passes through the training set), learning to classify each input as one of a fixed set of possible output catagories.
True

4. List some significant differences between the way that humans learn about objects and the way that ConvNets learn about objects.
ConvNets learn about a set number of categories of objects, whereas the types of objets human children learn is open-ended. In addition, children more actively explore and learn about the world, where ConvNets passively study the information given to them.

5. Why is it inaccurate to say that today's successful ConvNets "learn on their own?"
In order to create a ConvNet, humans have to do a great deal of work, such as gathering and labelling data and setting the parameters of the ConvNet.

6. In answer to the rhetorical question "Where does all of the data come from to fuel big data applications?," MM answers "You - and probably everyone you know." Please elaborate on the answer.
The data sets used to train big data applications are commonly scraped from the internet, often without the knowledge of those who took or uploaded the images. Uploading a picture to Facebook or Flickr means it can end up in one of these data sets.

7. How do car companies acquire the big data (labelled images of pedestrians, cyclists and other obstacles) needed to train robo-cars?
This data comes from cameras mounted on cars driving around.

8. What is the "long tail" phenomenon, and how does it relate to machines that learn (ConvNets)?

The long-tail phenomenon is the idea that if we could order every possible situation and AI could be faced with in order of probability, there would be an incredibly long tail of low-probability situations. Taken together, the probability of one of these events happening increases.

9. TRUE/FALSE - A commonly proposed solution to the long tail problem in AI systems is to complement supervised learning with unsupervised learning.
True

10. What is "unsupervised learning?"
This refers to a broad group of learning methods that do not involve labelled data.

11. What colorful remark did Yann LeCun make about unsupervised learning?
LeCun said that "unsupervised learning is the dark master of AI".

12. TRUE/FALSE - For general AI, almost all learning will have to be unsupervised, but no one has yet come up with the kinds of algorithms needed to perform successful unsupervised learning.
True

13. TRUE/FALSE - Humans have a fundamental competence lacking in current AI systems common sense. We have vast background knowledge of the world, both its physical and social aspects. We have a good sense of how objects - both animate and living - are likely to behave, and we use this knowledge extensively in making decisions about how to act in any given situation.
True

14. TRUE/FALSE - Many people believe that until AI systems have common sense as humans do, we won't be able to trust them to be fully autonomous in complex real-world situations.
True

15. TRUE/FALSE - Superficial changes to images, such as slightly blurring or speckling an image, changing some colors, or rotating objects in the scene, can cause ConvNets to make significant errors even when these perturbations don't affect humans' recognition of objects. This unexpected fragility of ConvNets – even those that have been said to "surpass humans at object recognition"' – indicates that they are overfitting to their training data and learning something different from what we are trying to teach them, a phenomenon that results in various manifestations of unreliability.
True

16. The unreliability of ConvNets can result in embarrassing – and potentially damaging – errors. Select a particularly embarrassing/damaging example of unreliability in ConvNets, and describe it in just a sentence or two.

In 2015, a ConvNet working as part of Google Photos labeled a photo of two black people as "gorillas". This exposed the racial bias inherent in this system and resulted in a public relations nightmare for Google.

17. At the end of the section on biased AI, MM observes that the problem of bias in applications of AI has been getting a lot of attention recently, with many articles, workshops, and even academic

research institutes devoted to this topic. What questions does she raise in conjunction with this observation? What do you think are the appropriate answers to these questions?

Mitchell asks if the data used to train AI should be as biased as our own society, or should it be tailored towards social reform aims. I would argue that it depends on the AI in question. An AI used in determining prisoner sentencing, for instance, if trained on racially biased legal decisions would perpetuate racial inequality in the legal system. However, if the point of the AI is to highlight discrepancies, to hold a mirror up to our society, unaltered data would be most effective.

18. TRUE/FALSE - You can often trust that people know what they are doing if they can explain to you how they arrived at an answer or a decision. However, "showing their work" is something that deep neural networks – the bedrock of AI systems – cannot easily do.

True

19. TRUE/FALSE - Recall that a convolutional neural network decides what object is contained in an input image by performing a sequence of mathematical operations (convolutions) propagated through many layers. For a reasonably sized network, these can amount to billions of arithmetic operations. While it would be easy to program the computer to print out a list of all the additions and multiplications performed by a network for a given input, such a list would give us humans zero insight into how the network arrived at its answer. A list of a billion operations is not an explanation that a human can understand.

True

20. What, according to MIT's Technology Review is the dark secret at the heart of AI?

This refers to the fact that neural networks cannot explain their reasoning in a way humans can understand.

21. What does the phrase "theory of mind" refer to, and how is it related to our interactions with AI systems such as deep networks?
Theory of mind refers to the notions we maintain about other humans, such as their motivations and knowledge base. Since humans don't usually have a theory of mind in relation to AI, it makes it hard to trust AI.

22. One of the hottest new areas of AI is variously called "explainable AI," "transparent AI," or "interpretable machine learning." To what do these terms refer?

These refer to neural networks that can show their work, so to speak. They show how they drew their conclusions in a way humans can understand.

23. The field of "adversarial learning" has emerged in response to the fact that AI systems can readily be fooled in dramatic fashion, like mixing up a guy in glasses with Milla Jovovich,

or misclassifying a stop sign for a speed-limit sign. Briefly describe the field of adversarial learning.

The field of adversarial learning centers around developing methods to fool various AI systems. A low-tech example of this is protesters wearing geometrically inspired face paint to confuse police facial recognition systems. A more technologically advanced example is slightly modifying a digital image so it looks identical to a human but causes the AI to categorize the image incorrectly.

24. Jeff Clune, an AI researcher at the University of Wyoming, made a very provocative analogy when he noted that there is "a lot of interest in whether Deep Learning is 'real intelligence' or a 'Clever Hans.'" Explain the essential question that underlies this analogy, being sure to incorporate a few words on the actual Clever Hans.

Clever Hans was a horse that appeared to be capable of arithmetic, but was really picking up on unconscious clues from its human to give the correct answer to the question. Similarly, it's unclear if AI is learning things like image recognition the way a human would understand it, or if it's just found clever tricks in the datasets to get the approved answer.