**COGNITIVE INTERLUDE**

We are coming to the end of our brief look into machine learning. Our focus was mostly on neural networks and backpropagation, but I remind the reader that there is much much more to ML than this. In addition, there are so-called *genetic algorithms* – and more broadly *adaptive computation* that includes all of the above. But we will set all this aside and now offer a high-level look at where we are and where we will go next.

Recall the notion of a KBN – knowledge base network – that was introduced earlier. This was a basic aspect of a cognitive agent – a system that makes decisions about what to do based on what it knows. And that knowledge can come in multiple forms. This in turn raised two questions: how is knowledge acquired, and how is it used? We have been exploring one answer to the first question: knowledge can come from a training process. But there are other answers as well. Knowledge can be gained by cultural transmission (especially language), and it can be gained by internal inference (reasoning). It is reasoning that we will turn to next.

But let us pause to consider where we are. It was from neuroscience (neuronal signaling) and cognitive psychology (Hebbian learning) that artificial neural networks (ANNs) arose. Now that the latter are big-time successes, can we look back and make conclusions about human learning? This is not so clear. While there have been a few suggestions as to how, for instance, backpropagation might conceivably be related to how the brain might learn via a form of training, this is generally thought to be unlikely. Nevertheless, ANNs continue to be a powerful inspiration for neuroscience, and vice versa.

One aspect of our study emphasized the use of randomly-chosen minibatches of training data, taken as statistical samples representative of the entire training set. When gradient descent is used with minibatches (or in the extreme case with only a single randomly-chosen training pair), it is referred to as *stochastic gradient descent*, or *SGD*. Such a statistical character may seem biologically implausible, but there is work indicating that many brain functions are like this, finely tuned to probabilities in the environment.

Note that a trained nnet can in some sense be said to have knowledge, in that it now can perform up to some standard of behavior: it has an ability or "know-how", that it lacked before training. For instance, it might now be able to classify inputs into types (digits 0-9, etc). We might say it can "recognize" handwritten digits. The same can be done for photographs of faces; a nnet can be trained to classify them as JFK, Marilyn Monroe, etc. It is tempting to say the nnet stores a "memory" of these faces.

But here we run the risk pointed out by Drew McDermott (in a famous paper, Artificial Intelligence Meets Natrual Stupidity), of letting a too-causal use of words mislead us. And as deep-learning pioneer Yoshua Bengio has said, "the machine is so stupid" – while others have noted that instead of worrying that machines are about

to get smarter than us and then take over the world, we should worry that they are incredibly stupid and already have taken over (much of) the world.

So we should be careful when saying that a system knows or remembers or recognizes this or that. It's ok as long as we know the technical meaning behind this, but it is easy to lose sight of. Will we someday be able to build machines that have human-like knowledge? Perhaps; it's too early to say. But it seems clear that nnet training *by itself* is insufficient. A system trained by traditional SGD does not retain, for instance, a record of its progress, cannot tell that it is now different from before, cannot even distinguish between this case of classifying a photo and that case (even of the same photo). It has no recognition of change or time. Could these be built in? Possibly; in fact, using recurrent (rather than feedforward) nnets people have done some amazing things along those lines.

But we are still at very early stages in this endeavor. It appears that some key linkages are needed among the different forms of knowledge, so that a system can know not only *how* to do X, but also *that* it can do X. This would seem to be essential for there to be useful decision-making. An agent should not decide to do things that it cannot do, but should consider things it can do (or can learn to do). This then bears on yet another form of knowledge: meta-knowledge (knowledge about oneself), a topic that will come up a little later.

In particular, it has been forcefully argued (by Hector Levesque and others) that finely tuned statistical training will not be enough to achieve human-level behavior. One major success area in statistical ML is natural-language processing. People have found that by applying training to millions of sentences found on the web, patterns can be detected that then allow systems to guess appropriate classifications of word-sequences with surprising accuracy. You may have noticed that over the past year or two, voice-driven menus have gotten much much better, from extremely annoying five years ago to mostly quite tolerable now. Machine translation has also gotten much better in that same time-frame (as in Google-Translate, for similar reasons.

On the other hand, whether such successes will eventually achieve human-level "understanding" is quite another matter. Levesque for instance argues that mere statistical pattern recognition is not up to the task of solving the *Winograd Schema Challenge*. Here is one example:
*The iron ball crashed through the table because it was made of light plywood.*
*What does "it" refer to?*
Levesque's claim is that a very large amount of *knowledge-that* is needed to resolve this. One needs to know lots about iron and wood and crashing and thinness and being "made of", and much more. Consider this alternate version:
*The ball crashed through the iron table because it was made of light plywood.*
This seems nonsensical. Levesque's point is that no amount to statistical patterning is adequate here; the meaning is exquisitely sensitive to the fine details not only of the words but of the world they refer to. It is as if (or maybe really is that) one

really envisions or imagines such a ball and table and their violent interaction in a sort of internal simulation. Yes, that surely *would* be largely based on past experience of patterns, for iron and wood and so on. But none of that determines what "it" refers to, without being able to identify specific items (this ball, that table, current crashing, etc) that the words refer to and how those meanings affect the envision interaction. And that seems to lead again to representations of time, memory, and reasoning.

Memory is sometimes spoken of a bit disparagingly in in computer science. That is, one hears that "memory is cheap" and one should not be overly concerned about it. But memory may play another role in addition to mere long-term storage. There is a famous book, Mind of a Mnemonist, by Russian neuroscientist Alexander Luria. Luria studied a patient who could not forget, and this was a tremendous handicap. His mind was constantly bombarded with impressions of a vast quantity of information, so that it was nearly impossible for him to function. It seems that memory goes hand in hand with forgetting (at least temporarily) so that one can focus on things of interest. And AI systems likely will have to have the same ability (to be selective about what information they are processing).

Psychologists have identified numerous forms of (human) memory. Consider the Wason experiment again. The subject has to look at the cards (or envelopes, etc), relate them to the rule that is being tested, decide whether more information is needed for a given card, remember that this decision has been made and that some other decisions still need to be made, and so on, all while *not* deciding what to eat for dinner or whether breakfast was satisfactory, etc. A robot that cannot direct its attention to certain elements of knowledge and not to others will be unlikely to have effective behavior for long.

I am using attention and memory a bit sloppily here. Let me give just a little more detail. Among the many forms of memory are these:

Short-term memory (STM)
Working memory
Long-term memory (LTM)
Semantic memory
Episodic memory

These are not intended to be entirely distinct. For instance, your memory of where you lived at age ten is surely a mixture of three of these at least. It is long-term, because your brain has kept it for a long time; it is semantic because this (e.g., you lived in California) is factual information that you understand in much the same way you would understand a similar fact about someone else; and it is episodic because (presumably) you actually recall some of the "episodes" (house, weather, friends, school, etc) you had experienced back then.

What about short-term and working memory? STM refers to the capacity to recall events or information without effort for up to roughly a minute or so, even if attention is distracted by something else. The exact time is not what matters here; it varies from person to person and event to event. But it is clear that our brains are able to recall fairly well certain things briefly after being presented, and then less and less so; while certain other things are often remembered for years. (There is also so-called intermediate-term memory, but we won't go into that here.)

Working memory refers to information being actively processed; for instance in the Wason experiment the subject must retain information about the task and what stage they are at and what they are trying to do at the moment. (So it is also episodic.) But – unlike Luria's mnemonist – we usually don't keep that sort of information around once the task is done; we retain it as long as it is being used.

So STM and working memory sound different, and are defined differently. But in practice it can be hard to tease them apart, and it has been suggested that they might really be the same thing. That is, short-term memory might be just what is needed to keep things "around" for awhile as one attends to a task. The fact that not every retained item is relevant to the task just might mean that the mechanisms cannot always perfectly determine what is and is not relevant, since tasks can vary so greatly. In any event, note that "what I am doing now" is central to working memory, and thus the latter seems to involve some sort of self-notion. Meta-knowledge appears over and over in cognition, and will reappear later on for us, more than once.