GRESHAM COLLEGE
19 MARCH 2019

DEEP LEARNING: MIRACLE OR SNAKE OIL?

PROFESSOR RICHARD HARVEY FBCS

The full story of artificial intelligence has yet to be told but it includes panic, hype, tragedy and comedy. The story includes huge triumphs in the face of ludicrous expectations, and spectacular failures in the face of very reasonable expectations. As I write this, we are at one of the periodic highs in the "hype cycle" that appears to drive AI. And at the peaks of such cycles there are large amounts of time and money invested, so it would be wrong to puncture the balloon of AI expectations. That said, some modest weighing of the pros and cons is certainly permissible. Of course there are many critics of AI, but many of those critics appear to be woefully ignorant of how the AI machines work which makes them easy pickings for the technocrats on Wittgensteinian grounds[1].

I'd like to focus on one type of AI: classification. Classification is by no means the only activity in AI and it is not the most controversial, but it has the great advantage that it is easy to understand and there is a clear progression of ideas since the 1950s. I'd further like to specialise to talk about artificial neural network classifiers (ANNs). ANNs are not the only fruit of machine learning research, and there are many people, me included, who cross the street to avoid ANNs, but they are definitely part of the hype cycle and, the current craze "deep" learning was motivated by ANNs.

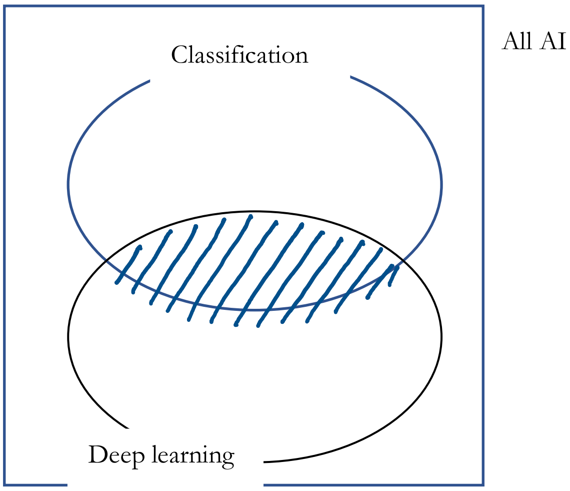

Figure 1: *Venn diagram of AI showing the topic of the lecture (the hatched region).*

Interest in neural inspired machines started with McCullogh and Pitts [1] who constructed mathematical models of neurons. They postulated that a neuron had essentially two components: a summer which weighted the inputs and then a non-linearity which modelled the neuron firing. It is essentially a straw on a camel's back model – once the sum exceeds a threshold, then the camel reacts. In the first model, called the Perceptron, Rosenblatt [2] modelled the nonlinearity as "all or nothing" — the perceptron outputs +1 if the weighted sum of the inputs is greater than zero otherwise -1. These models were exciting because, firstly, Rosenblatt actually built one – its

[1] "Whereof one cannot speak, thereof one must be silent".

about the size of eight washing machines – and secondly because it was shown that perceptions could compute logical AND and OR functions. This latter property meant that classical AI, which at that time was concerned with logical reasoning, could be modelled by connectionist AI so, to paraphrase the bar-room conversations of the time, it was only a matter of time before we built an electronic brain. There are several PhDs yet to be written about humankind's fascination with building a human brain. Suffice it to say, if one is ever incautious enough to riff on the possibility of replicating human intelligence in front of a journalist, there will be hell to pay the next day.

For perceptrons, a critical intervention was the publication of a book [3][2]. In *Perceptrons*, Minsky and Papert noted a number of deficiencies with perceptrons. One, which has passed into folklore, is that a Perceptron was unable to model the logical EXOR function. With the benefit of hindsight one might cry "So what?" But as is often the case with a hype cycle, it only takes modest criticism for the balloon to deflate very rapidly. To overcome the EXOR problem ones needs layers of perceptrons (many neurons). Such a grouping is called an Artificial Neural Network or ANN. A fundamentally tricky part of ANN was how to train them and, when Minsky and Papert wrote their book, ANNs looked untrainable hence work came to a stop. To explain the solution to this we need a slight digression to explain supervised learning.

In supervised learning we have a collection of patterns and their desired class. The role of the machine is to learn a mapping from the patterns (or features) to the classes. To pick a simple example we could imagine the problem of determining someone's sex by examining their height and shoe-size. Here, sex is the class (which we might represent with +1 for male and -1 or female) and f = [shoe-size, height] is the feature vector. There is a whole literature on how to code the classes into codes that are helpful for classifiers and also on how to choose and normalise the features.

With neural classifiers the basic idea is to guess some random weights and then to measure the classification. Of course many points are misclassified, but by considering them one at a time, we can use the classification error to adjust the weights. Early versions of the perceptron used a hard clipping function so that the classifier was either right or wrong (+1 or -1). The great problem with such systems, as in life, is that they allow minimal feedback. An innovation was to replace the hard clipping function with a smoother version called a sigmoid. The sigmoid allows big errors to make bigger adjustments to the weights than the smaller ones — we can propagate the error back through the network, hence *backpropagation*. This new learning algorithm, backpropagation, or backprop, was the key to the resurgence of ANNs in the 1980s. There was much excitement about the capabilities of these new machines. But there were significant criticisms of these too. One of the main ones was that the early networks were rather unprincipled – it was not clear what the output scores meant nor was it clear how they had been derived. This was fairly soon sorted out [4] and it became clear that ANNs could be trained to give outputs that estimated the probabilities. However there were continuing doubts about the blackbox element of the artificial neural network that, coupled with the fact that there were other classifiers that were equally effective, led to a loss of popularity of ANNs. They became untrendy.

A further difficulty was that, as the network complexity was increased to handle more challenging classification problems, the training became intractable. It seemed that backprop was not making enough changes to the earlier network layers and the training stalled. We had entered another "AI winter" and one of the long-term proponents of neural networks, Geoff Hinton [5], later described how he used to have to downplay the fact that he was using neural networks in order to get his work published. The numbers of people working in neural networks diminished from thousands to dozens, but Hinton and a few acolytes realised that if they could overcome the training difficulties with deeper networks, then there was potential. This new form is what is known as deep learning[3]. These new networks have been spectacularly successful and all of us working in pattern recognition now have difficulty getting papers accepted unless they are using deep learning!

[2] This was highly unusual since Science does not use books for publishing new results — books are humanities scholars and undergraduates — rather the article is the preferred form.

[3] Deep learning is not a well-defined term. It is a portmanteau phrase developed to cover a library of methods that are needed to train deep networks. Indeed there is some evidence that the term was invested to sneak neural network papers past scientific reviewers who were known to be antagonistic to ANNs.

Despite their effectiveness, there remains considerable anxiety about the use of deep learning. These appear to stem principally from the difficulty of really knowing how a network makes its decisions[4]. Of course this lack of transparency is frustrating to scientists because one of the attractions of pattern recognition is the prospect of a better understanding of the physics of particular situations[5]. However it also has some serious consequences. If we do not know how an algorithm makes it decisions, what confidence can we have that it is fair or unbiased? [6]. This leads to new concerns — we need to devise ways to explicitly test for bias in algorithms. These methods are closer to the types of tests we might use for monitoring bias in humans and represent a new direction for AI.

A final and intriguing aspect to deep learning is that it has been rapidly taken up by commercial providers. Microsoft, IBM, Apple, Google, Facebook, Amazon, and many many others have large teams of people working on deep learning. The resources available to these commercial teams far exceed those available to universities so many of the worlds most effective machine learning systems are protected by commercial secrecy — some details are published and some are not. A further bizarre consequence has been that, as part of their evangelism for deep learning, several commercial providers have made their deep learning software easily available. It is so simple to use that a high-school student can easily download and implement a deep network. Deep networks have another novel property which is transferability — I can download, say, Alexnet which is one of the well known image recognition networks and retrain it for my task — the lower levels of the deep network barely change (they took weeks of learning from huge databases so that is very nice to know) and only a modest amount of data is needed to train the higher levels. When you conjoin the easy availability of pre-trained networks, with simple ways of retraining networks with easy accessibility to non-experts, we have the situation where the number of people building deep learning systems considerably exceeds the number of people who know how they work.

Recently a large computer manufacturer interviewed people for a machine learning position. They short-listed forty applicants, all with PhDs in machine learning, but only one of their shortlist was able to answer basic theoretical questions about how the network was trained. It seems that we are building very effective systems that no-one understands nor really wants to understand. Depending on your viewpoint, that is either a triumph of democratisation of a science that was formerly the preserve of only a few experts, or a moment of considerable danger. It is notable that a number of large IT companies who are active in machine learning have very recently hired ethicists or have publicly made available work in bias or fairness in AI. Either way I believe that it represents a unique position in the history of computer science and it will be interesting to see how it turns out.

1. McCulloch, W and Pitts, W. (1943), A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, 5:115-133.
2. Speech of Hon Hugh L Carey, Tribute to Dr Frank Rosenblatt, Congressional Record, Proceedings and Debates of the 92nd Congress, First Session, US Government Printing Office, 1971.
3. Minsky, M. and Papert, S.,1972 (2nd edition with corrections, first edition 1969) Perceptrons: An Introduction to Computational Geometry, The MIT Press, Cambridge M
4. Bishop, C M, (1986), Neural Networks for Pattern Recognition, Oxford University Press
5. https://www.thestar.com/news/world/2015/04/17/how-a-toronto-professors-research-revolutionized-artificial-intelligence.html
6. Buolamwini, J. and Gebru, T.,(2018), Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, Proceedings of the 1st Conference on Fairness, Accountability and Transparency, pp 77–99, 81, Proceedings of Machine Learning Research, http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf,

[4] Although I should point out that given there is a very large programme of work on visualising deep networks and on transparent learning we should remain optimistic of progress in this area.

[5] My own work considers computer lip-reading. For me it is nice that we have produced one of the best systems in the world but that is not the prize — the prize is knowing what aspects of the problem are difficult and why.