Knowledge and Truth in Machine Learning

Ioannis Votsis (Northeastern University London)

One key question in the epistemology of science is to what extent scientific theories/models provide any knowledge of the world. Another way of asking, more or less, the same thing focuses on the extent to which assertions made by such theories/models are veridical or verisimilar. The recent successes enjoyed by machine learning (ML), and particularly deep learning (DL), in detecting patterns, fitting functions and extracting features raises corresponding questions about the use of such models in science. To what extent do they encode knowledge of the world? To what extent do they make assertions that are veridical or verisimilar? This talk attempts to provide partial answers to these questions, with one eye on the unique circumstances and details that are characteristic of DL models, namely the black box nature of their representations and the peculiar role of simplicity considerations in DL model selection and construction.

Let us be clear from the outset that the scope of this talk is restricted to scientific applications of DL. That is, we target DL model building and assessment in the service of science. By this we mean that the data employed for training is of scientific interest, but also the trained models are putative scientific models. As such, it's pertinent to ask questions about the explanatory merits, empirical adequacy and truthlikeness of those models.

Now, it is well known that DL models are capable of reaching heights of accuracy, the likes of which had not been seen before in attempts to automate science. It is also well known that DL models are exceedingly complex. One, admittedly extreme, case involves the latest large language models. Briganti (2024), for example, reports that GPT-4o is composed of 175 billion parameters (also known as weights). Roughly speaking, the more complex a model, the less we should expect to understand its representations of the world. The trouble lies in the highly distributed nature of those representations, as their constituent elements are spread out over countless nodes and weights, making the task of deciphering how they represent very hard; some would even say, impossible.

To reduce the complexity of DL models, a number of methods and techniques are employed. One promising method, or class of methods, is regularisation. Its main aim is to reduce overfitting (avoid models with 'high variance'), e.g. by penalising large weights. Regularisation methods include: L1, L2, dropout, early stopping, batch normalisation, and data augmentation. Another promising method is pruning. The idea goes back at least to Le Cun, Denker and Solla's 'Optimal Brain Damage', where they assert: "We introduce a new technique called Optimal Brain Damage (OBD) for reducing the size of a learning network by selectively deleting weights… The basic idea of OBD is that it is possible to take a perfectly reasonable network, delete half (or more) of the weights and wind up with a network that works just as well, or better." (1989: 598). How does one decide which parameters to delete? Well, in Le Cun et al.'s proposal they suggest "deleting parameters with small 'saliency', i.e. those whose deletion will have the least effect on the training error" (599). There are in fact two types of pruning: structured and unstructured. The former involves removing whole structures, i.e. nodes and weights, from deep neural nets (DNNs). The latter involves

removing just the weights.[1] Lots of developments have been made in pruning since that paper was published. In recent years, for example, it has been shown that pruning can lead to models that are "less than 10-20% of the [original model's] size", and that enjoy comparable or even increased levels of accuracy (Frankle and Carbin 2018). To summarise, pruning reduces complexity by attempting to remove elements that play no positive role in a network's accuracy. The main upshots include fewer training examples, decreased computation demands, decreased memory demands and the same or even better generalisability.

There are different ways of understanding what goes on in such cases. On the one hand, it may be argued that a DL model, despite being highly accurate (read: successful), is merely a stopgap (read: not truthlike), and hence does not offer any genuine knowledge of the world. In other words, its success is instrumental and not a reliable indicator of truth-likeness, i.e. its ability to faithfully represent the world. On this view, pruning is important for purely pragmatic reasons, e.g. we prune DL models to increase the training speed of models, and to decrease their computation and memory demands. One the other hand, it may be argued that a highly accurate DL model cannot be completely devoid of truth content. After all, if at least some of its outputs correspond to independent measurements of the world, that suggests that its representation, however convoluted, encodes some aspects of the world's structure. Moreover, if a pruned model not only maintains earlier levels of accuracy, but also builds on them (read: increased success), this suggests that simpler models are likely to be closer to the truth. This chimes well with the epistemically positive claim that we shouldn't be overly concerned with the failures or idleness of some elements in a theory or model.

This talk will explore how these two readings of what goes on in DL modelling and pruning can potentially be reconciled. It will also investigate what this reconciliation means for the possibility of determining to what extent, and under what conditions, DL model representations can be veridical or verisimilar.

References:

Briganti, G. (2024). How ChatGPT works: a mini review. *European Archives of Oto-Rhino-Laryngology*, 281(3), 1565-1569.

Frankle, J., & Carbin, M. (2018). 'The lottery ticket hypothesis: Finding sparse, trainable neural networks'. arXiv preprint arXiv:1803.03635.

LeCun, Y., Denker, J., & Solla, S. (1989). Optimal brain damage. *Advances in neural information processing systems*, vol. 2.

[1] Pruning can also be applied to other types of AI modelling, e.g. symbolic decision tree modelling where tree branches are removed instead, but we ignore these here.