Machine Learning, Interpretability and the Scientific Realism Debate

Ioannis Votsis (Northeastern University London)

The main point of contention in the scientific realism debate is whether successful scientific theories or models reveal truths about the unobservable world. Scientific realists (Vickers 2022) answer in the affirmative while anti-realists (Wray 2018) answer in the negative. One area of disagreement is the role played by simplicity and, more generally, explanatory power. Scientific realists insist that theories or models with increased simplicity or explanatory power are more likely to be successful and truthlike. Scientific anti-realists deny this connection, claiming instead that increased simplicity or explanatory power is a pragmatic feature of theory/model choice. In this talk, we consider what, if anything, machine learning (ML) can tell us about this dispute. More precisely, we examine two powerful ML techniques, namely pruning and the application of integrated gradients, whose aim is to simplify/increase the interpretability of models, and ask how these techniques affect the debate over scientific realism.

As is well known, deep learning (DL) models may be capable of enjoying high degrees of accuracy but are often complex (e.g. containing hundreds of neurons) and explanatorily impermeable. What is perhaps less well-known is that there are various ML techniques that aim to reduce the size of DL models and increase their interpretability. One promising approach is *pruning* and involves the selective removal of parameters from models, which are then trained/retrained faster using the original datasets (Le Cun, Denker and Solla 1989). The upshot is 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). Another promising approach involves *integrated gradients*. This allows the differential evaluation of the contributions made by inputs to outputs, thereby helping us understand and visualise (Qi, Khorram and Li 2019) the relative importance of features in a model.

These developments provide some reasons, which prima facie seem to favour the anti-realist, and others, which prima facie seem to favour the realist. Take the anti-realist reasons first. That a model can be highly accurate (and thus successful) but merely a stopgap (and thus not truthlike) suggests that success is not a reliable indicator of truth-likeness. Moreover, since one of the motivating factors for pruned (and thus simpler) models is faster training, this suggests a pragmatic role for simplicity in ML. Now take the realist reasons. That ML models can be substantially simplified without loss of accuracy/success suggests that their pre-treatment (i.e. before pruning and the application of integrated gradients) counterparts may possess a kernel of truth within them, even though it is buried under unnecessary layers of complexity. Furthermore, that a post-treatment model can achieve higher accuracy (and thus success) suggests that removing those unnecessary layers of complexity may be tantamount to something like de-idealisation and hence increased closeness to the truth. The talk will explore how these reasons can potentially be reconciled and whether the subsequent reconciliation provides an advantage for one or the other side of the debate.

References:

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.

Qi, Z., Khorram, S., & Li, F. (2019). Visualizing Deep Networks by Optimizing with Integrated Gradients. In CVPR workshops (Vol. 2, pp. 1-4).

Vickers, P. (2022). *Identifying Future-Proof Science*. Oxford University Press.

Wray, K. B. (2018). *Resisting scientific realism*. Cambridge University Press.