Can an artificial neural network teach us quantum physics?

Anna Dawid

Machines beat human masters in chess and Go, and even win one-to-one in Dota 2. How do they do it? Of course, they have an advantage of huge memory, but while it enables to store all possible moves in chess, there are more potential moves in Go than there are atoms in the universe, so without something 'more' a machine could win with a world number one in Go just by accident. Programmers and scientists are to be praised (or blamed), who, inspired by how a human brain works, found a way to teach a machine how to learn. Creating such algorithms is what machine learning is all about. Such learning machines, besides hurting gamers' pride, filter our spam, recognise voices and faces, even drive cars. They have not started to conquer the world yet.

They help not only in everyday life. When there is a need to analyse an enormous amount of data and find hidden patterns or to find a very complicated function basing on a few accurate but expensive measurements or calculations, scientists use machine learning algorithms as well, especially in quantum chemistry, material science and biology. Quantum many-body physics since always has struggled with a problem of a huge amount of data caused by the exponential growth of a complexity of a wave function describing a system along with a number of objects contained in it as well as due to a great number of nontrivial correlations between these objects. It is no wonder then that the machine learning algorithms found their use also here. For example, it turned out that they can classify phases in physical systems pretty well. There is, however, a dangerous trait that all learning machines possess. The more complex they are, the less understandable for humans they become. And this increase of complexity does not always go in pair with the improvement of its performance. For example, it was shown that state-of-the-art neural networks can be easily fooled into recognising a panda as a gibbon with an imperceptibly (for a human) little noise. It is one of many examples when neural networks do not work exactly as we supposed they did, and which show we should not fully trust their predictions. It is astonishing that nonetheless learning machines are used worldwide. As Pedro Domingos stated, '*People worry that the computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world*'. A growing number of people finds this lack of understanding how complex neural networks work crucial in everyday life, and addressing this problem became the job of a booming sub-field of machine learning, called the machine learning interpretability, that design methods aiming to discover the internal logic of learning machines.

Analogous problems hold for neural networks used in physical problems. Even though they classify phases at much lower computational cost than more conventional methods, they only recover known phase diagrams, and basically, have not taught scientist anything new about quantum physics so far. Even more, no one can really be sure that machines learn anything corresponding to known theories of phase transitions like an order parameter. We will make first steps towards understanding what neural networks used in phase classification problems really learn, and we will do that with help of interpretability methods such as influence functions and heat maps. When applied correctly, they can indicate the most influential training examples for a specific machine's prediction, or find the parts of data that are the most discriminative for a chosen class.

The results of this project can be threefold. Firstly, they can provide the first strong indication, or even proof, that machines used in chosen phase classification problems do indeed learn an order parameter. However, it can also be discovered that neural networks follow nothing correlated to this parameter. In such a case, they either learn some noise, background, other information unconsciously provided by a researcher, or, what would be especially exciting, they learn something relevant for physics, different than order parameter, but not noticed before. Consequences for the field will be of great importance: we will either validate the use of machine learning methods in quantum physics, discredit it, or truly provide it with a power of teaching quantum physicists something new.