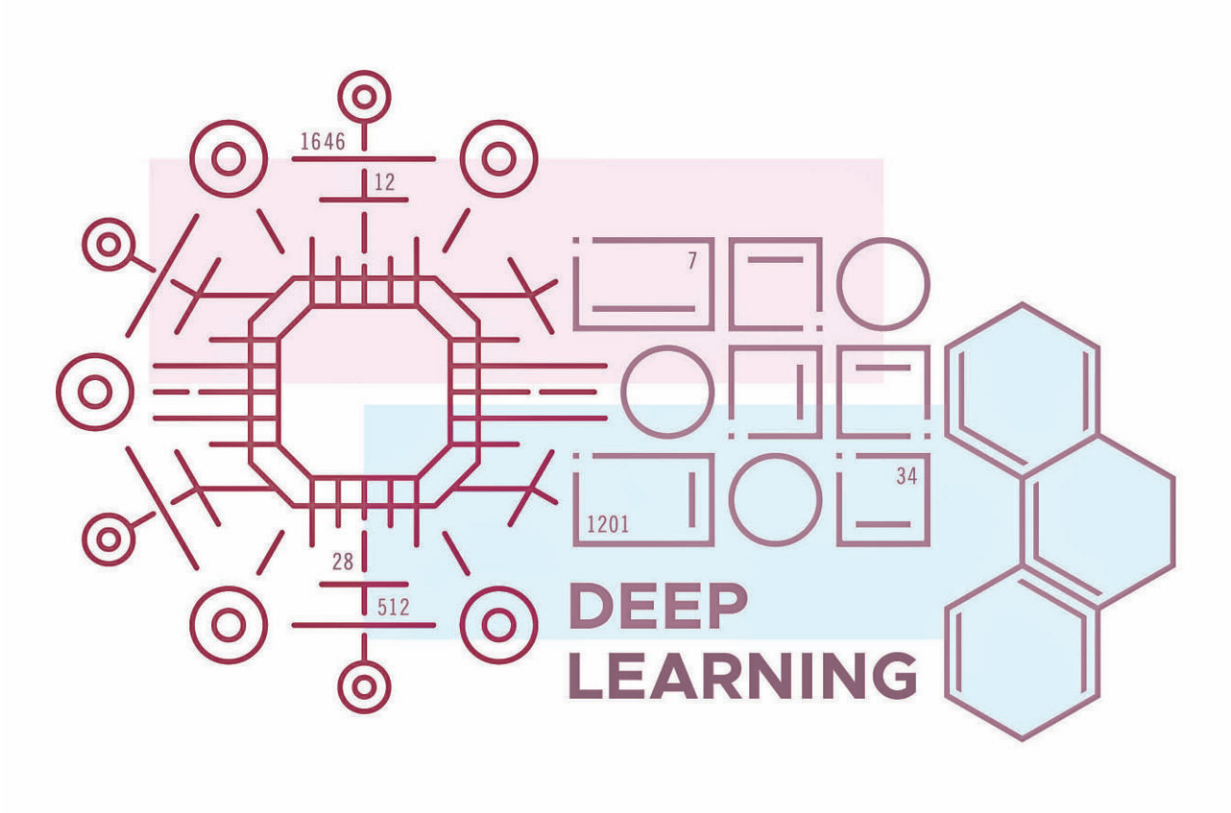
New AI approach bridges the 'slim-data gap' that can stymie deep learning approaches

February 26 2019, by Tom Rickey



PNNL's deep learning network tackles tough chemistry problems with the aid of some pre-training. Credit: Timothy Holland/PNNL

Scientists have developed a deep neural network that sidesteps a problem that has bedeviled efforts to apply artificial intelligence to tackle

complex chemistry—a shortage of precisely labeled chemical data. The new method gives scientists an additional tool to apply deep learning to explore drug discovery, new materials for manufacturing, and a swath of other applications.

Predicting chemical properties and reactions among millions upon millions of compounds is one of the most daunting tasks that scientists face. There is no source of complete information from which a deep learning program could draw upon. Usually, such a shortage of a vast amount of clean data is a show-stopper for a deep learning project.

Scientists at the Department of Energy's Pacific Northwest National Laboratory discovered a way around the problem. They created a pre-training system, kind of a fast-track tutorial where they equip the program with some basic information about chemistry, equip it to learn from its experiences, then challenge the program with huge datasets.

The work was presented at KDD2018, the Conference on Knowledge Discovery and Data Mining, in London.

Cats, dogs, and clean data

For deep learning networks, abundant and clear data has long been the key to success. In the cat vs. dog dialogue that peppers discussions of AI systems, researchers recognize the importance of "labeled data—a photo of a cat is marked a cat, a dog is marked a dog, and so on. Having many, many photos of cats and dogs, clearly marked as such, is a good example of the type of data that AI scientists like to have. The photos provide clear data points that a neural network can use to learn from as it begins to differentiate cats from dogs.

But chemistry is more complex than sorting cats from dogs. Hundreds of factors affect a molecule's promiscuity, and thousands of interactions

can happen in a flash of a second. AI researchers in chemistry are often faced with either small but thorough data sets or huge but inconsistent datasets—think 100 clear images of chihuahuas or 10 million images of furry blobs. Neither is ideal or even workable alone.

So the scientists created a way to bridge the gap, combining the best of "slim but good data" with "big but poor data."

The team, led by former PNNL scientist Garrett Goh, employed a technique known as rule-based supervised learning. Scientists point the neural network to a vast repository of chemical data known as ChEMBL, and they generate rule-based labels for each of these many molecules, for example calculating the mass of the molecule. The neural network crunches through the raw data, learning principles of chemistry that relate the molecule to basic chemical fingerprints. Taking the neural network trained on the rule-based data, the scientists presented it with the small, but high quality, dataset containing the final properties to be predicted.

The pre-training paid off. The program, called ChemNet, achieved a level of knowledge and precision as accurate or more than the current best deep learning models available when analyzing molecules for their toxicity, their level of biochemical activity related to HIV, and their level of a chemical process known as solvation. The program did so with much less labeled data than its counterparts and achieved the results with less computation, which translates to faster performance.

More information: Garrett B. Goh et al. Using Rule-Based Labels for Weak Supervised Learning: A ChemNet for Transferable Chemical Property Prediction. arXiv:1712.02734 [stat.ML]. arxiv.org/abs/1712.02734

Provided by Pacific Northwest National Laboratory