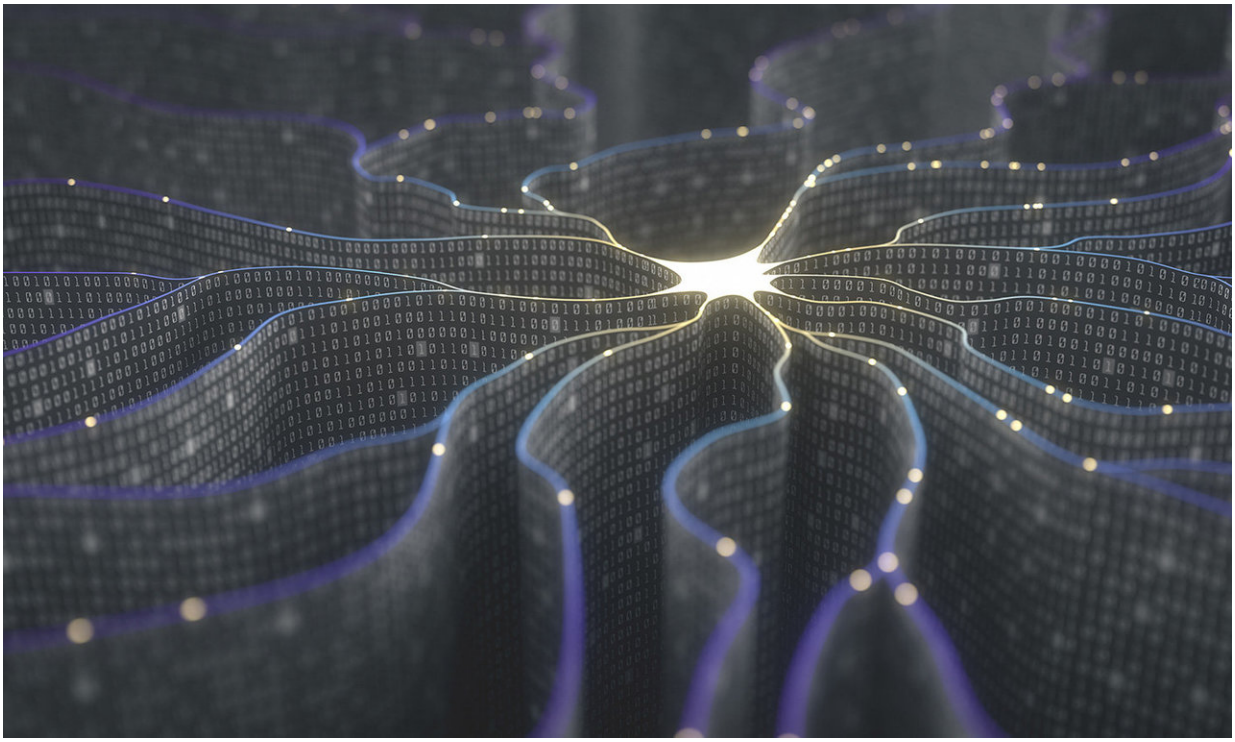
Algorithm leverages Titan supercomputer to create high-performing deep neural networks

November 29 2017



Inspired by the brain's web of neurons, deep neural networks consist of thousands or millions of simple computational units. Leveraging the GPU computing power of the Cray XK7 Titan, ORNL researchers were able to auto-generate custom neural networks for science problems in a matter of hours as opposed to the months needed using conventional methods. Credit: Oak Ridge National Laboratory

Deep neural networks—a form of artificial intelligence—have demonstrated mastery of tasks once thought uniquely human. Their triumphs have ranged from identifying animals in images, to recognizing human speech, to winning complex strategy games, among other successes.

Now, researchers are eager to apply this computational technique—commonly referred to as deep learning—to some of science's most persistent mysteries. But because scientific data often looks much different from the data used for animal photos and speech, developing the right artificial neural network can feel like an impossible guessing game for nonexperts. To expand the benefits of deep learning for science, researchers need new tools to build high-performing neural networks that don't require specialized knowledge.

Using the Titan supercomputer, a research team led by Robert Patton of the US Department of Energy's(DOE's) Oak Ridge National Laboratory (ORNL) has developed an evolutionary algorithm capable of generating custom neural networks that match or exceed the performance of handcrafted artificial intelligence systems. Better yet, by leveraging the GPU computing power of the Cray XK7 Titan—the leadership-class machine managed by the Oak Ridge Leadership Computing Facility, a DOE Office of Science User Facility at ORNL—these auto-generated networks can be produced quickly, in a matter of hours as opposed to the months needed using conventional methods.

The research team's algorithm, called MENNDL (Multinode Evolutionary Neural Networks for Deep Learning), is designed to evaluate, evolve, and optimize neural networks for unique datasets. Scaled across Titan's 18,688 GPUs, MENNDL can test and train thousands of potential networks for a science problem simultaneously, eliminating poor performers and averaging high performers until an optimal network emerges. The process eliminates much of the time-

intensive, trial-and-error tuning traditionally required of machine learning experts.

"There's no clear set of instructions scientists can follow to tweak networks to work for their problem," said research scientist Steven Young, a member of ORNL's Nature Inspired Machine Learning team. "With MENNDL, they no longer have to worry about designing a network. Instead, the algorithm can quickly do that for them, while they focus on their data and ensuring the problem is well-posed."

Pinning down parameters

Inspired by the brain's web of neurons, deep neural networks are a relatively old concept in neuroscience and computing, first popularized by two University of Chicago researchers in the 1940s. But because of limits in computing power, it wasn't until recently that researchers had success in training machines to independently interpret data.

Today's neural networks can consist of thousands or millions of simple computational units—the "neurons"—arranged in stacked layers, like the rows of figures spaced across a foosball table. During one common form of training, a network is assigned a task (e.g., to find photos with cats) and fed a set of labeled data (e.g., photos of cats and photos without cats). As the network pushes the data through each successive layer, it makes correlations between visual patterns and predefined labels, assigning values to specific features (e.g., whiskers and paws). These values contribute to the weights that define the network's model parameters. During training, the weights are continually adjusted until the final output matches the targeted goal. Once the network learns to perform from training data, it can then be tested against unlabeled data.

Although many parameters of a neural network are determined during the training process, initial model configurations must be set manually.

These starting points, known as hyperparameters, include variables like the order, type, and number of layers in a network.

Finding the optimal set of hyperparameters can be the key to efficiently applying deep learning to an unusual dataset. "You have to experimentally adjust these parameters because there's no book you can look in and say, 'These are exactly what your hyperparameters should be,'" Young said. "What we did is use this evolutionary algorithm on Titan to find the best hyperparameters for varying types of datasets."

Unlocking that potential, however, required some creative software engineering by Patton's team. MENNDL homes in on a neural network's optimal hyperparameters by assigning a neural network to each Titan node. The team designed MENNDL to use a deep learning framework called Caffe to carry out the computation, relying on the parallel computing Message Passing Interface standard to divide and distribute data among nodes. As Titan works through individual networks, new data is fed to the system's nodes asynchronously, meaning once a node completes a task, it's quickly assigned a new task independent of the other nodes' status. This ensures that the 27-petaflop Titan stays busy combing through possible configurations.

"Designing the algorithm to really work at that scale was one of the challenges," Young said. "To really leverage the machine, we set up MENNDL to generate a queue of individual networks to send to the nodes for evaluation as soon as computing power becomes available."

To demonstrate MENNDL's versatility, the team applied the algorithm to several datasets, training networks to identify sub-cellular structures for medical research, classify satellite images with clouds, and categorize high-energy physics data. The results matched or exceeded the performance of networks designed by experts.

Networking neutrinos

One science domain in which MENNDL is already proving its value is neutrino physics. Neutrinos, ghost-like particles that pass through your body at a rate of trillions per second, could play a major role in explaining the formation of the early universe and the nature of matter—if only scientists knew more about them.

Large detectors at DOE's Fermi National Accelerator Laboratory (Fermilab) use high-intensity beams to study elusive neutrino reactions with ordinary matter. The devices capture a large sample of neutrino interactions that can be transformed into basic images through a process called "reconstruction." Like a slow-motion replay at a sporting event, these reconstructions can help physicists better understand neutrino behavior.

"They almost look like a picture of the interaction," said Gabriel Perdue, an associate scientist at Fermilab.

Perdue leads an effort to integrate neural networks into the classification and analysis of detector data. The work could improve the efficiency of some measurements, help physicists understand how certain they can be about their analyses, and lead to new avenues of inquiry.

Teaming up with Patton's team under a 2016 Director's Discretionary application on Titan, Fermilab researchers produced a competitive classification network in support of a neutrino scattering experiment called MINERvA (Main Injector Experiment for v-A). The task, known as vertex reconstruction, required a network to analyze images and precisely identify the location where neutrinos interact with the detector—a challenge for events that produce many particles.

In only 24 hours, MENNDL produced optimized networks that

outperformed handcrafted networks—an achievement that would have taken months for Fermilab researchers. To identify the high-performing network, MENNDL evaluated approximately 500,000 neural networks. The training data consisted of 800,000 images of neutrino events, steadily processed on 18,000 of Titan's nodes.

"You need something like MENNDL to explore this effectively infinite space of possible networks, but you want to do it efficiently," Perdue said. "What Titan does is bring the time to solution down to something practical."

Having recently been awarded another allocation under the Advanced Scientific Computing Research Leadership Computing Challenge program, Perdue's team is building off its deep learning success by applying MENDDL to additional high-energy physics datasets to generate optimized algorithms. In addition to improved physics measurements, the results could provide insight into how and why machines learn.

"We're just getting started," Perdue said. "I think we'll learn really interesting things about how deep learning works, and we'll also have better networks to do our physics. The reason we're going through all this work is because we're getting better performance, and there's real potential to get more."

AI meets exascale

When Titan debuted 5 years ago, its GPU-accelerated architecture boosted traditional modeling and simulation to new levels of detail. Since then, GPUs, which excel at carrying out hundreds of calculations simultaneously, have become the go-to processor for deep learning. That fortuitous development made Titan a powerful tool for exploring artificial intelligence at supercomputer scales.

With the OLCF's next leadership-class system, Summit, set to come online in 2018, deep learning researchers expect to take this blossoming technology even further. Summit builds on the GPU revolution pioneered by Titan and is expected to deliver more than five times the performance of its predecessor. The IBM system will contain more than 27,000 of Nvidia's newest Volta GPUs in addition to more than 9,000 IBM Power9 CPUs. Furthermore, because deep learning requires less mathematical precision than other types of scientific computing, Summit could potentially deliver exascale-level performance for deep learning problems—the equivalent of a billion billion calculations per second.

"That means we'll be able to evaluate larger networks much faster and evolve many more generations of networks in less time," Young said.

In addition to preparing for new hardware, Patton's team continues to develop MENNDL and explore other types of experimental techniques, including neuromorphic computing, another biologically inspired computing concept.

"One thing we're looking at going forward is evolving deep learning networks from stacked layers to graphs of layers that can split and then merge later," Young said. "These networks with branches excel at analyzing things at multiple scales, such as a closeup photograph in comparison to a wide-angle shot. When you have 20,000 GPUs available, you can actually start to think about a problem like that."

More information: Steven R. Young et al. Evolving Deep Networks Using HPC, *Proceedings of the Machine Learning on HPC Environments - MLHPC'17* (2017). DOI: 10.1145/3146347.3146355

Adam M. Terwilliger et al. Vertex reconstruction of neutrino interactions using deep learning, *2017 International Joint Conference on Neural Networks (IJCNN)* (2017). DOI: 10.1109/IJCNN.2017.7966131

Provided by Oak Ridge National Laboratory