Science | DOI:10.1145/2771283 Chris Edwards

Growing Pains for Deep Learning

Neural networks, which support online image search and speech recognition, eventually will drive more advanced services.

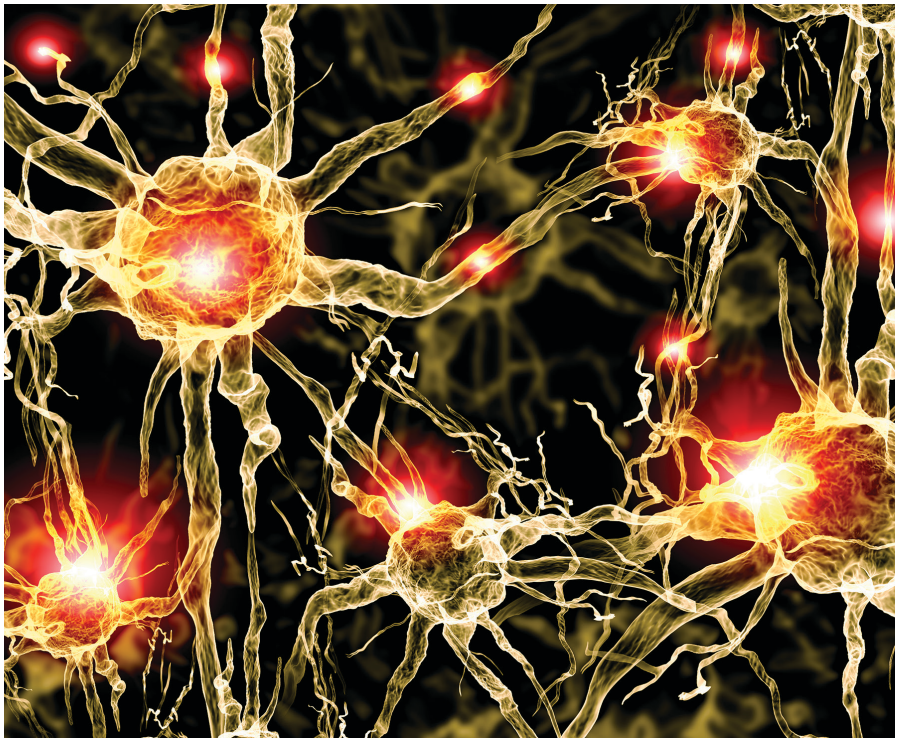

IMAGE BY SERGEY NIVENS

Advances in theory and computer hardware have allowed neural networks to become a core part of online services such as Microsoft's Bing, driving their image-search and speech-recognition systems. The companies offering such capabilities are looking to the technology to drive more advanced services in the future, as they scale up the neural networks to deal with more sophisticated problems.

It has taken time for neural networks, initially conceived 50 years ago, to become accepted parts of information technology applications. After a flurry of interest in the 1990s, supported in part by the development of highly specialized integrated circuits designed to overcome their poor performance on conventional computers, neural networks were outperformed by other algorithms, such as support vector machines in image processing and Gaussian models in speech recognition.

Older simple neural networks use only up to three layers, split into an input layer, a middle 'hidden' layer, and an output layer. The neurons are highly interconnected across layers. Each neuron feeds its output to each of the neurons in the following layer. The networks are trained by iteratively adjusting the weights that each neuron applies to its input data to try to minimize the error between the output of the entire network and the desired result.

Although neuroscience suggested the human brain has a deeper architecture involving a number of hidden layers, the results from early experiments on these types of systems were worse than for shallow networks. In 2006, work on deep architectures received a significant boost from work by

Geoffrey Hinton and Ruslan Salakhutdinov at the University of Toronto. They developed training techniques that were more effective for training networks with multiple hidden layers. One of the techniques was 'pre-training' to adjust the output of each layer independently before moving on to trying to optimize the network's output as a whole. The approach made it possible for the upper layers to extract high-level features that could be used more efficiently to classify data by the lower, hidden layers.

Even with improvements in training, scale presents a problem for deep learning. The need to fully interconnect neurons, particularly in the upper layers, requires immense compute power. The first layer for an image-processing application may need to analyze a million pixels. The number of connections in the multiple layers of a deep network will be orders of magnitude greater. "There are billions and even hundreds of billions of connections that have to be processed for every image," says Dan Cireşan, researcher at the Manno, Switzerland-based Dalle Molle Institute for Artificial Intelligence Research (IDSIA). Training such a large network requires quadrillions of floating-point operations, he adds.

Researchers such as Cireşan found it was possible to use alternative computer architectures to massively speed up processing. Graphics processing units (GPUs) made by companies such as AMD and nVidia provide the ability to perform hundreds of floating-point operations in parallel. Previous attempts to speed up neural-network training revolved around clusters of workstations that are slower, but which were easier to program. In one experiment in which a deep neural network was trained to look for characteristic visual features of biological cell division, Cireşan says the training phase could have taken five months on a conventional CPU; "it took three days on a GPU."

Yann LeCun, director of artificial intelligence research at Facebook and founding director of New York University's Center for Data Science, says, "Before, neural networks were not breaking records for recognizing continuous speech; they were not big enough. When people replaced Gaussian models with deep neural nets, the error rates went way down."

Deep neural nets showed an improvement of more than a third, cutting error rates on speech recognition with little background noise from 35% to less than 25%, with optimizations allowing further improvements since their introduction.

There are limitations to this form of learning. London-based DeepMind—which was bought by Google in early 2014 for $400 million—used computer games to evaluate the performance of deep neural networks on different types of problems. Google researcher Volodymyr Mnih says the system cannot deal with situations such as traversing a maze, where the rewards only come after successfully completing a number of stages. In these cases, the network has very little to learn from when it tries various random initial maneuvers but fails. The deep neural network fares much better at games such as Breakout and Virtual Pinball, where success may be delayed, but it can learn from random responses.

When it comes to deploying deep networks in commercial applications, teams have turned to custom computer designs using field-programmable gate arrays (FPGAs). These implement custom electronic circuits using a combination of programmable logic lookup tables, hard-wired arithmetic logic units optimized for digital signal processing, and a matrix of memory cells to define how all of these elements are connected.

Chinese search-engine and web-services company Baidu, which uses deep neural networks to provide speech recognition, image searches, and to serve contextual advertisements, decided to use FPGAs rather than GPUs in production servers. According to Jian Ouyang, senior architect at Baidu, although individual GPUs provide peak floating-point performance, in the deep neural network applications used by Baidu, the FPGA consumes less power for the same level of performance and could be mounted on a server blade, powered solely from the PCI Express bus connections available on the motherboard. A key advantage of the FPGA is that because the results from one calculation can be fed directly to the next

ACM Member News

ELLIOTT CONSIDERS ONE MORE BIG PROJECT

Chip Elliott has a triple-threat career: he is chief scientist at Raytheon BBN Technologies, adjunct professor of computer science at Dartmouth College, and Futures Director at the National Science Foundation's Global Environment for Network Innovations (GENI). Elliott is also an active inventor, with over 90 issued patents.

Elliott was born in Connecticut, raised in Kansas, and "escaped back to New England" to earn a B.A. in mathematics from Dartmouth College in New Hampshire. He began his computer science career as a programmer, teaching and founding the startup True Basic. His passion is research "seven to 10 years beyond current economic drivers, to explore [my vision]."

One such project involved collaborating with quantum physicists and photonics experts to build the world's first quantum network, linking Raytheon BBN with Harvard and Boston universities.

His research at GENI centers on large-scale cloud infrastructures. GENI provides 3,000 researchers at 50 U.S.-based universities with a virtual laboratory to promote innovation. "Most clouds are pre-baked; we bake our own," Elliott says, adding, "Our researchers are programming new types of clouds that are better at processing big data."

Among the initiatives he is keen to pursue at BBN are "futuristic cellular Internets"; fully homomorphic encryption—"done correctly, you can perform computation on encrypted data without unencrypting it"; synthetic biology, "to make cells into engineered systems, performing advanced tasks like producing diesel fuel inside plant cells," and producing machines that "sense interactions with chemicals, light, and pressure, to incorporate in the human bloodstream" for the potential medical benefits.

—Laura DiDio

without needing to be held temporarily in main memory, the memory bandwidth requirement is far lower than with GPU or CPU implementations.

"With the FPGA, we don't have to modify the server design and environment, so it is easy to deploy on a large scale. We need many functions to be supported that are impossible to deploy at the same time in FPGA. But we can use their reconfigurability to move functions in and out of the FPGA as needed. The reconfiguration time is less than 10µs," says Ouyang.

The Baidu team made further space savings by using a simplified floating-point engine. "Standard floating-point implementations provided by processors can handle all possible exceptions. But in our situation we don't need to handle all of the exceptions of the IEEE [754] standard."

As well as finding ways to use more effective processors, researchers are trying to use distributed processing to build more extensive deep-learning networks that can cope with much larger datasets. The latency of transfers over a network badly affects the speed of training. However, rearranging the training algorithms together with a shift from Ethernet networking to Infiniband, which offers lower latency, allowed a team from Stanford University in 2013 to achieve almost linear speed-ups for multiple parallel GPUs. In more recent work using clusters of CPUs rather than GPUs, Microsoft developed a way to relax the synchronization requirements of training to allow execution across thousands of machines.

More scalable networks have made it possible for Baidu to implement an "end to end" speech recognition system called Deep Speech. The system does not rely on the output of traditional speech-processing algorithms, such as the use of hidden Markov models to boost its performance on noisy inputs. It reduced errors on word recognition to just over 19% on a noise-prone dataset, compared to 30.5% for the best commercial systems available at the end of 2014.

However, pre-processing data and combining results from multiple smaller networks can be more effective than relying purely on neural networks. Cireşan has used a combination of image distortions and

Researchers are trying to use distributed processing to build more extensive deep-learning networks that can cope with much larger datasets.

"committees" of smaller networks to reduce error rates compared to larger single deep-learning networks. In one test of traffic-sign recognition, the combination of techniques resulted in better performance than human observers.

Deciding on the distortions to use for a given class of patterns takes human intervention. Cireşan says it would be very difficult to have networks self-learn the best combination of distortions, but that it is typically an easy decision for humans to make when setting up the system.

One potential issue with conventional deep learning is access to data, says Neil Lawrence, a professor of machine learning in the computer science department of the University of Sheffield. He says deep models tend to perform well in situations where the datasets are well characterized and can be trained on a large amount of appropriately labeled data. "However, one of the domains that inspires me is clinical data, where this isn't the case. In clinical data, most people haven't had most clinical tests applied to them most of the time. Also, clinical tests evolve, as do the diseases that affect patients. This is an example of 'massively missing data.'"

Lawrence and others have suggested the use of layers of Gaussian processes, which use probability theory, in place of neural networks, to provide effective learning on smaller datasets, and for applications in which the neural networks do not perform well, such as data that is interconnected across many different databases, which is the case in healthcare. Because data may not be present in certain databases for a given candidate, a probabilistic model can deal with the situation better than traditional machine-learning techniques. The work lags behind that on neural networks, but researchers have started work on effective training techniques, as well as scaling up processing to work on platforms such as multi-GPU machines.

"We carry an additional algorithmic burden, that of propagating the uncertainty around the network," Lawrence says. "This is where the algorithmic problems begin, but is also where we've had most of the breakthroughs."

According to Lawrence, deep-learning systems based on Gaussian processes are likely to demand greater compute performance, but the systems are able to automatically determine how many layers are needed within the network, which is not currently possible with systems based on neural networks. "This type of structural learning is very exciting, and was one of the original motivations for considering these models."

In currently more widespread neural-network systems, Cireşan says work is in progress to remove further limitations to building larger, more effective models, "But I would say that what we would like mostly is to have a better understanding of why deep learning works." C

Further Reading

Hinton, G.E., and Salakhutdinov, R.R.
Reducing the dimensionality of data with neural networks, *Science* (2006), Vol 313, p 504.

Schmidhuber, J.
Deep learning in neural networks: an overview, *Neural Networks* (2015), Volume 61, pp85-117 (ArXiv preprint: http://arxiv.org/pdf/1404.7828.pdf)

Mnih, V., et al
Human-level control through deep reinforcement learning, *Nature* (2015), 518, pp529-533

Damianou A.C. and Lawrence N.D.
Deep Gaussian processes, *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS) 2013.* (ArXiv preprint: http://arxiv.org/pdf/1211.0358.pdf)

Chris Edwards is a Surrey, U.K.-based writer who reports on electronics, IT, and synthetic biology.