

Accelerating the Power of Deep Learning With Neural Networks and GPUs

AI goes beyond image recognition.

Abstract

Deep learning using neural networks and graphics processing units (GPUs) is starting to surpass machine learning for image recognition and other applications. Deep neural networks are helping to advance self-driving cars, faster development of new drugs, and real-time multiple-language translation for online chats.

Find out more about:

- Machine learning versus deep learning
- The power of inference
- GPUs delivering near real-time performance

The Next Stage in Machine Learning

Deep learning using neural networks has sparked a revolution in many areas of advanced machine learning and perception. Its potential for use in machine vision was demonstrated conclusively in 2012 during the annual ImageNet challenge, where research teams competed with programs designed to classify and detect things for the ImageNet visual database. A convolutional neural network called AlexNet outperformed all the conventional machine vision competitors by a substantial margin.

Since then, deep neural networks have performed exceedingly well across various algorithms originally developed for machine learning-based image recognition. Now, deep neural networks are expanding the potential uses well beyond image recognition, from self-driving cars to quicker development of new drugs to real-time translation in several languages for online chats.

Accelerated deep learning dramatically enhances the training process for AI systems. Faster and better inference capabilities improve the performance of Internet of Things applications, including sensor data analytics. Sensors generate large volumes of data that can, with sufficient training, be used to predict machine malfunctions, for example. But traditional machine learning methods can take months. Accelerated deep learning techniques, including automatic pattern recognition and reinforcement learning, decrease the time required to train an AI system—and therefore decrease the time to truly useful inference capabilities.

Now, deep neural networks are expanding the potential uses well beyond image recognition, from self-driving cars to quicker development of new drugs to real-time translation in several languages for online chats.

Machine Learning v. Deep Learning

Figure 1 depicts a high-level view of the process flow for traditional machine learning approaches: raw data (i.e., images) are analyzed (feature extraction) and then identified, classified, or detected (depending on the goal of the application), and then a result is delivered to the user.

The problem with this approach is that it relies heavily on human intervention for the feature extraction, identification, classification, and detection phases.



Figure 1: Traditional machine learning approach to visual perception.

Deep learning approaches require much less human effort in designing learning parameters—deep learning applications essentially learn how to learn—and this means they offer exciting opportunities and problem solving capabilities.

Deep learning approaches require much less human effort in designing learning parameters—deep learning applications essentially learn how to learn—and this means they offer exciting opportunities and problem solving capabilities.

Deep Learning	Traditional Machine Learning
Architecture driven	Features driven
Automatic	Manual
Deep	Shallow
Lots of data	More technique
Compute intensive	Human intensive
Automates learning	Expertise and domain barrier

Inference Makes the Difference

While traditional machine learning applications are trained to deliver accurate results, deep learning neural network applications are trained—and then taught to make inferences to handle a broader class of inputs and deliver constantly improving results.

After forward propagation is completed, the results are compared against a set of well-understood correct answers to validate and compute for error data. The backward propagation stage sends errors back through the network's layers and updates their weights using a gradient descent algorithm. This helps the process improve its performance to the appropriate level. For example, if an application needs to be right 99.9% of the time, the learning process will be more rigorous than if it only needs to be right 80% of the time.

Figure 2 (next page) shows a schematized view of the training and deployment phases of a deep neural network approach. Training a deep neural network involves designing parameters that include examples of inputs and suggested outputs. Training can last several hours or several weeks, depending on the complexity of the task, and uses forward and backward propagation.

After training, the deep learning neural network is deployed to run inference computations using its previously learned parameters to classify, recognize, and process unknown inputs.

Inference is extremely useful for classifying images, localizing faces, or real-time speech translation.

In the inference phase, the system learns how to learn so that its outputs get better and better over time (and in real-time).

For inference, the performance goals are different from those associated with training. To minimize the network's end-to-end response time, inference typically batches a smaller number of inputs than would be used for training. Use cases relying on inference—for example, machine vision for an autonomous car—are required to be as responsive as possible and to reduce wait times.

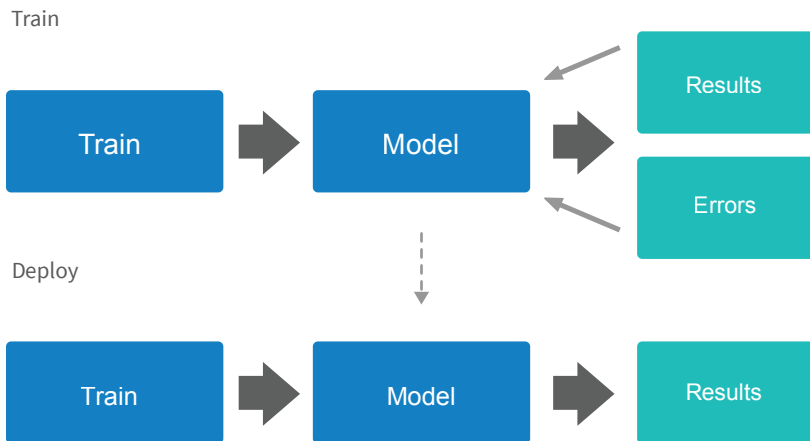


Figure 2: Deep learning applications are trained using forward and backward propagation, but the system continues to improve in the deployment stage when applications are taught to make inference.

Use cases relying on inference—for example, machine vision for an autonomous car—are required to be as responsive as possible and to reduce wait times.

In general, training involves a higher workload than inference (see Figure 3). During training, the most important factor is high throughput in terms of data volume, but during inference, speed and accuracy of results become paramount. This necessitates a kind of balancing act for designers of deep learning neural networks.

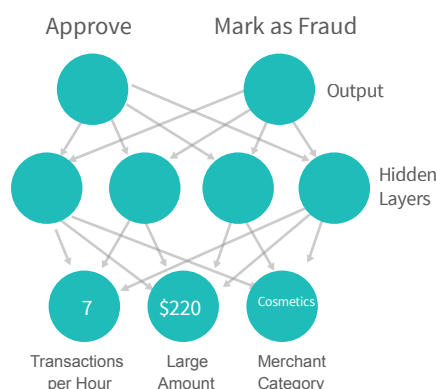
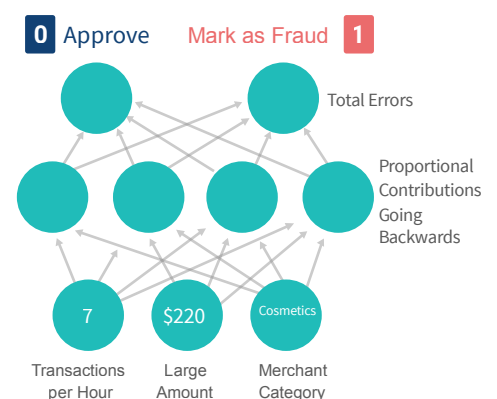



Figure 3: Deep learning training compared to inference: In training, many inputs, often in large batches, are used to train a deep neural network.

Figure 3 (continued): In inference, the trained network is used to discover information within new inputs that are fed through the network in smaller batches.





It's possible to batch thousands of inputs for training and operate on them simultaneously during training of deep neural networks.

Most deep learning applications strive to maximize a batch size while keeping latency under a given application-specific maximum. This enables inference computations to occur at various scales, all the way from cloud applications running in a large-scale datacenter to real-time inference, such as pedestrian detection on an embedded processor inside an autonomous vehicle.

GPUs Excel at Neural Network Inference

Deep learning neural networks demand extraordinary processing power because deep learning involves a lot of vector and matrix operations. Using customized deep learning cores can lead to stellar performance for neural network processing implementations. However, for developing new networks, it is also valuable to be able to work within a standard framework that makes the process of testing and modifying networks fast and lightweight.

Graphics processing units (GPUs), designed for 3D computer graphics and 60 fps for game play, offer outstanding performance for deep learning applications. While both CPUs and GPUs can handle graphical operations, GPUs accelerate graphical computations very well. GPUs perform faster because of the distributed/parallel nature of the architecture with many low-end processing nodes.

Deep learning algorithms run several times faster on a GPU compared to a CPU, and learning times can be reduced from months to weeks or even a day. It's possible to batch thousands of inputs for training and operate on them simultaneously during training of deep neural networks to prevent overfitting (which can occur when a model is too complex) and manage loading weights from GPU memory across several inputs, increasing computational efficiency.

This acceleration is important because researchers or people working with deep learning will want to experiment with multiple deep learning architectures—for example, the number of layers, cost functions, regularization methods, and others—to discover exactly what works best for their deep learning application.

The Bottom Line

Deep learning is fast emerging as a key component of a wide variety of challenging and incredibly useful artificial intelligence applications, and GPUs offer a powerful tool for exploring the power and potential of deep learning. These tools represent a true paradigm shift for the machine learning endeavor, both in terms of how we approach the challenge of teaching computers to think and how we use thinking computers in the real world.

About The Author**Sunil Karkera****Head of Digital Reimagination™ Studio**

Sunil is an engineer by profession, and over the past 20 years, he founded three successful startups in Silicon Valley (wireless products, SaaS solutions, and mobile applications), led information systems groups at News Corporation, and worked and consulted for companies such as Nest, Westinghouse, Dell-Sonicwall, Lockheed, and Siebel. He is also a trained typographer.

About Artificial Intelligence at TCS and our Digital Reimagination™ Studio

At TCS, we operate with the belief that the future of business will be driven by five powerful digital forces: mobility and pervasive computing, the cloud, Big Data, social media, and artificial intelligence (AI). We are applying AI, notably deep learning, to all kinds of applications from autonomous vehicles to the analysis of sensor data from the Internet of Things, from fraud detection to natural language processing and conversational agents.

The TCS Digital Reimagination™ Studio is dedicated to helping businesses create fundamentally new experiences by reimagining industries through creative thinking. The Studio brings a start-up culture to large enterprise clients by leveraging the best of world-class creative, design, engineering, and business domain experts. The result is business transformation through rapid product prototyping and extremely agile collaboration.

Contact

To learn more, contact the TCS Digital Reimagination™ Studio at analytics.insights@tcs.com

Subscribe to TCS White Papers

TCS.com RSS: http://www.tcs.com/rss_feeds/Pages/feed.aspx?f=w

Feedburner: <http://feeds2.feedburner.com/tcswhitepapers>

About Tata Consultancy Services Ltd (TCS)

Tata Consultancy Services is an IT services, consulting, and business solutions organization that delivers real results to global business, ensuring a level of certainty no other firm can match. TCS offers a consulting-led, integrated portfolio of IT and IT-enabled infrastructure, engineering, and assurance services. This is delivered through its unique Global Network Delivery Model™, recognized as the benchmark of excellence in software development. A part of the Tata Group, India's largest industrial conglomerate, TCS has a global footprint and is listed on the National Stock Exchange and Bombay Stock Exchange in India.

For more information, visit us at www.tcs.com