WHITEPAPER

Equipping Enterprise Data Science for Deep Learning: What IT Leaders Need to Know

Deep learning is a form of artificial intelligence that utilizes neural networks, which are computing systems inspired by the human brain and nervous system — essentially a multi-layered "mesh" architecture. Neural networks are not new, but their use in tackling machine learning problems has become so specialized and valuable, it has emerged as the discipline of deep learning.

As with machine learning (ML), a data scientist engaging in deep learning (DL) uses models to predict the future, reveal hidden information, identify structure in large data sets, and find unusual trends and patterns in data. The magic of DL models is in how well they handle data with a huge number of input variables and/or very complex relationships between input variables.

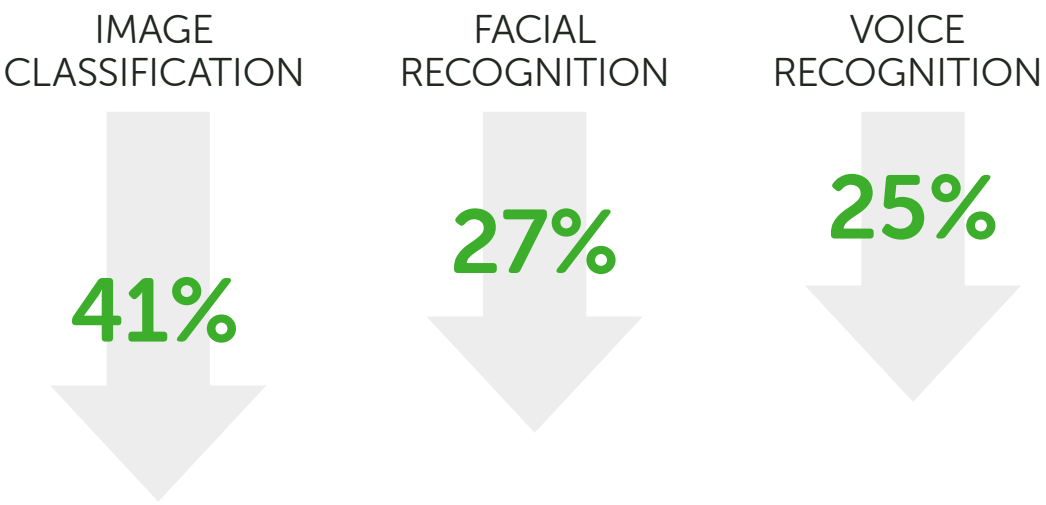
McKinsey & Company estimates that deep learning has the potential to create up to $5.8 trillion in value annually across nine business functions in 19 industries. No wonder so many enterprise organizations are starting to sit up and take notice.

Performance: Deep learning vs. machine learning

When the number of input variables and the complexity of relationships between them are very great, deep learning techniques outperform traditional machine learning. This is often the case with image classification, natural language processing, and complex anomaly detection. For example, a relatively common DL model for image classification takes as input 150,000 values (per image!) and predicts one of 20,000 image categories. This would be extremely hard to handle with other ML techniques.

DEEP LEARNING OFTEN OUTPERFORMS MACHINE LEARNING

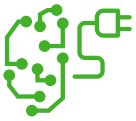
REDUCTION IN ERROR RATE ACHIEVED BY DEEP LEARNING VS. TRADITIONAL MODELS



Source: McKinsey and Company

Typical deep learning applications

IMAGE CLASSIFICATION
is being used to improve radiological diagnostics in healthcare, replace scanners in retail stores, identify types of damage in homeowner and automobile insurance reports, and more.



NATURAL LANGUAGE PROCESSING (NLP)
makes it possible for a search engine to suggest terms when you're typing in the search box, and enables virtual assistants to understand what you're saying and translate it into terms the computer can process.

COMPLEX ANOMALY DETECTION
is used to identify exceptional patterns in vast data sets — used in financial fraud detection but also for manufacturing and other process-oriented business to uncover hidden areas for improvements.

These applications are even more valuable to businesses when used in combination. For example, image recognition and anomaly detection can enable a publisher to discover printing flaws before a book or magazine goes into binding. Combining NLP and image recognition makes it possible for airlines to leverage photographs and even mechanics' handwritten notes to improve maintenance performance.

There are other DL applications that don't have parallels in traditional machine learning. For example, DL transformation models can be used to remove noise from images, age/rejuvenate faces, or apply a specific painter's style to images. Generation is another example, in which a DL model is used to create new examples of things using past examples and a smaller number of inputs. Examples of generative models include voice synthesis and landscape image generation.

Historical obstacles to deep learning in the enterprise

Until quite recently, only behemoths like Amazon or Google could afford to implement deep learning at scale. Most enterprise organizations faced the usual IT obstacles — time and budget.

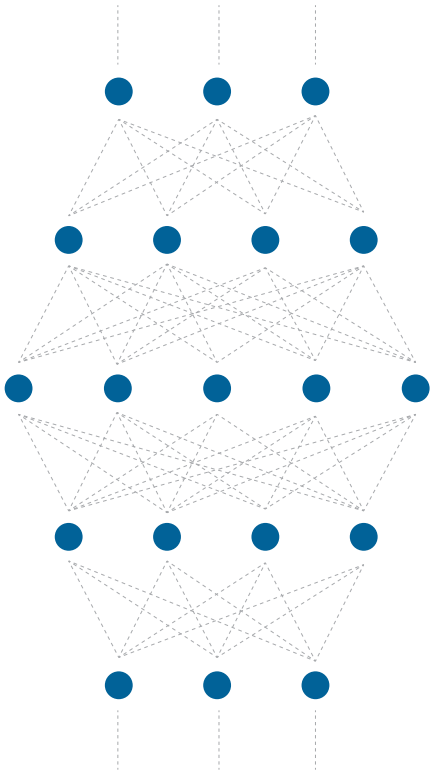
First, let's take a look at the issue of time. Assuming reasonable data sets, traditional machine learning models usually take seconds or minutes to train. Training a complex deep learning model once with unoptimized hardware can take days or even weeks. What's more, a model must be trained many times during development, as it is adjusted and improved. Most enterprise data science teams are under a certain amount of pressure to deliver something of value and, until relatively recently, deep learning just took too long.

Second, the start-up costs for deep learning were often prohibitive. To enable a data scientist to try deep learning on a particular business problem, IT would have to:

- Figure out how to upgrade an existing workstation with a USD $400-$3,000 GPU (which could also take skills and time they may not have).
- Purchase a dedicated GPU server or workstation at a cost of USD $5,000–$50,000 (depending on the number and type of GPUs).

This only makes sense if you're 100% sure of the long-term value and utilization of those machines.

These obstacles are even more significant when you factor in the reality that a lot of deep learning is exploration. A data scientist might set out to use DL techniques only to find out there isn't enough data to use it effectively, or that this approach simply doesn't fit the problem. Executive leadership might not be thrilled might not be thrilled with the idea of spending a whole lot of time and money on something that doesn't pan out.



New technology advances make deep learning more accessible

As data scientists have continued to explore and experiment with deep learning techniques, the demand for better technologies and tools has pushed both manufacturers and the open-source community to deliver. Neither science nor industry will stand for a cumbersome, time-consuming technique when it's possible to speed things up.

Here's how advances have stacked up to accelerate deep learning and — by doing so — democratize it for the enterprise and beyond:

1

FASTER, SPECIALIZED GPUS

Deep learning is built on massively parallel array calculations. Deployed as a co-processor, the GPU architecture is ideal for speeding these calculations up.
In response to data scientists' demands, GPU manufacturers have designed specific GPUs for deep learning applications. With a single GPU co-processor alongside the CPU, the DL model training process can be 10x faster (or more) than CPU-only training. For models that can be scaled up to use multiple GPUs, the training process can be accelerated even more.

2

GPU-OPTIMIZED SOFTWARE TOOLS

To take advantage of a GPU, the DL model has to be built for it. As data scientists began using GPUs more frequently, manufacturers and software developers responded with GPU-optimized APIs, deep learning frameworks and fundamental math libraries. This new world of GPU-optimized software tools is like a well-stocked hardware store, with everything data scientists need to work on their projects, right at their fingertips. Data scientists still have to do the work, but they no longer have to make their own tools and materials first.

3 OPEN-SOURCE DL PACKAGES

The open-source community is the source of tremendous innovation in deep learning, and in AI as a whole.

Now there's an entire stack of open-source DL tools, which include tutorial materials and a huge number of examples data scientists can learn from. Anyone can modify the packages to try out new ideas, without having to start from absolute zero. This saves a lot of time and effort.

What's more, researchers at the cutting edge of deep learning use the very same open-source software packages in their work. This greatly shortens the time for new discoveries to move from publication to practical use. For example, a paper about a new NLP approach for text classification is likely to be written using PyTorch or TensorFlow, tools that are already familiar to anyone wanting to use this new approach.

4 CLOUD GPU COMPUTING

The emergence of Cloud GPU offerings let data scientists get to work with GPU-accelerated deep learning right away — there's no waiting for the procurement process to work its way through enterprise approvals and no time lost to IT setting up systems.

Cloud GPUs can also make deep learning projects much more affordable. A data scientist or IT department can rent GPU servers in the cloud at a cost of USD $1-$25 per hour (depending on the number and kind of GPUs they need in one server). Because most data science projects don't require a GPU, it makes a lot of sense to pay for the GPU only when you need it. And being able to get started for $1/hr is very appealing.

There are a few caveats: Like all cloud services, cloud GPU can cost more than owning the hardware if you have extremely large data sets or expect very high utilization. In addition, some businesses have security policies that prohibit uploading data to the cloud, or lack a security plan that's designed for cloud computing.

How enterprise IT leaders can plan for GPU-accelerated deep learning

For deep learning projects, data scientists need access to a server or high-end workstation with a powerful CPU, plenty of memory, and a GPU co-processor. In some cases, they need access to more than one of these machines. However, no data science team runs exclusively on GPU computing. Buying one GPU or even an entire GPU workstation for every data scientist may be overkill, depending on your team and their needs.

Here are some guidelines for IT leaders looking to equip their teams for deep learning without overtaxing their budgets.

ESTIMATE YOUR CAPACITY NEEDS

Work with your data science team to arrive at a reasonable estimate of GPU usage. The best-practice recommendation is "only one user per GPU" — meaning that it's best to allow only one application or data scientist to use the GPU at a time. This gives the user the most GPU memory for their training batches and ensures maximum responsiveness. Note: If a single data scientist is training several different models at once, they will likely need one GPU per simultaneous training session. Additionally, deploying a given DL model may or may not require a GPU. Try to make your GPU capacity estimates as realistic as possible.

BUY ONLY WHAT YOU NEED

The most cost-efficient approach for supporting a deep learning practice is to implement a heterogeneous cluster with mixture of GPU and non-GPU nodes, such as CPUs. Keeping in mind that GPUs are commonly used in pairs, the following guidelines may help illustrate the number of GPUs needed for different use cases.

The sweet spot for most organizations is likely 2-6 GPUs per system, depending on the tasks and cost containment requirements. Each GPU includes 2-4 cores, and more cores are needed for projects that require extensive data processing before model training. CPU memory should be 2-3x GPU memory, or more if the training dataset is very large.

CONSIDER CLOUD GPU

It's a great way to get started, but know the trade-off point. If you know your expected utilization, you can compare the costs of on-premise versus the cloud and make a decision up front. If you can't forecast utilization accurately, keep an eye on the numbers as your deep learning practice ramps up. Once the cost of cloud outstrips the cost of owning your own GPU systems, it's time to pull things in house.

GPU COUNT PER SERVER

LOW (1-2)

Suitable for pilot studies and initial testing, but not very high density

MODERATE (3-7)

Good for cluster cost efficiency when typical user allocation is 1: GPU

HIGH (8-16)

Best for HPC and Specialized Deep Learning applications

Anaconda Enterprise makes it easy to explore the potential of deep learning

Anaconda Enterprise is an enterprise-ready, secure, and scalable data science platform that empowers teams to govern data science assets, collaborate, and deploy data science projects. Anaconda Enterprise provides a scalable foundation for the enterprise to explore the potential of deep learning by accessing open-source innovation and managing compute resources effectively.

World-class machine/deep learning requires petaflop-scale model training, made economically viable and more practical via GPUs and automated deployment into production IT environments. Anaconda Enterprise makes it easy for IT leaders to manage GPU resources and for data scientists to be more productive in deep learning projects. Users can simply check out a GPU when needed (e.g., for training a deep learning model). When the job completes, Anaconda Enterprise automatically returns the GPU to the cluster. This approach makes sharing GPUs across an organization cost-effective while also ensuring availability for users.

Anaconda has curated the most popular deep learning frameworks for Anaconda Enterprise and packed them with GPU acceleration to handle even complex and heavy DL workloads. Within Anaconda Enterprise, these DL frameworks can be combined with your data science team's favorite Python packages—including pandas, Dask, and Jupyter—to power data science experiments and production deployments.

To schedule a demo of Anaconda Enterprise or to speak to a sales rep, visit **anaconda.com/contact**.

ANACONDA.