ovum

Machine Learning in Business Use Cases

Artificial intelligence solutions that can be applied today

Publication Date: 20 Apr 2015 | Product code: IT0022-000335

Michael Azoff

Summary

Catalyst

The subfield in artificial intelligence (AI) known as machine learning or cognitive computing has in recent years become highly active as techniques such as deep learning and IBM's Watson initiative have given rise to improved performance in their generalization capabilities. Deep learning in particular has benefited from being ported to run on Nvidia's programmable graphics processing units (GPUs), reducing the neural network system training time from weeks to a day or less, and giving rise to the most sophisticated neural networks yet devised that have broken established benchmarks in image classification and speech recognition. As a result, there has been a surge of interest from startups to established businesses looking to exploit these new techniques. Large enterprises such as Google, Microsoft, and Facebook are investing in AI, and the large IT services and SI players are investing in robotic process automation (RPA) to transform the office workplace. Visible changes due to AI can be expected in society in the next five to 10 years.

Ovum view

The AI field is going through a new step change in capabilities, similar to Backpropagation, with deep learning systems offering improved discrimination capabilities when separating signal from noise. These systems are accelerated by running on Nvidia's new-generation programmable GPUs, reducing training time from weeks to a day or less in some cases. Typical deep learning applications cover image recognition (tracking a person in a crowd, for example), as well as speech recognition and understanding, including understanding in a first-time exposure to a voice (the system has not been trained to understanding only one person's speech pattern), a Holy Grail in AI. Current best accuracy is the 95% region using deep learning. According to Andrew Ng at Baidu, achieving 99% accuracy appears within reach and will transform human-machine interaction, with voice commands able to be distinguished by machines even in highly noisy environments. Today's 95% accuracy is already seeing business applications available on the market.

The use of robotics in manufacturing (albeit with limited intelligence) is a mature field, and RPA, which is technology that enables robots to learn applications in the business workplace and automate transaction processing and other data-related tasks, is the office equivalent that is already gaining early adoption. Robots exploiting advances in AI will also fuel growth in the burgeoning robotics industry, providing robots with sufficient intelligence to perform physical tasks. The future for robots is therefore particularly promising: robot domestic servants in the home (for cleaning, assisting elderly with heavy lifting tasks, and so on), robot workers in the office, and (more ominously if you have watched the Terminator films) robot soldiers. The new generation of robots is proving its usefulness in being able to navigate complex terrain encountered for the first time, and being able to interact with humans through speech and vision.

Because human-machine interfaces have not changed much since the invention of the computer mouse, there is likely to be a step change in how humans interact with machines with the improvements that deep learning introduces. AI has been quietly proving itself for many years in fields such as business intelligence with predictive analytics, application performance analytics, credit risk profiling and scoring, and fraud detection. The new generation of solutions exploiting deep learning

will enable humans to talk to software applications and robots, and these machines will have an intelligent understanding of their environment through visual observation.

The combination of the invention of deep learning algorithms, the acceleration in training these systems on GPUs, and the rapid dissemination of AI knowledge through an open community that shares papers, source code, and AI frameworks, has created a step change that will see AI advances permeate society in multiple ways.

Key messages

- Deep learning-based neural networks have created a step change improvement in the accuracy of AI systems, enabling human-machine interface through speech.
- Business applications powered by deep learning are now available.
- The combination of deep learning algorithms accelerated on GPUs has been the pivotal breakthrough that has also accelerated progress in the field.
- There are two kinds of AI: Type A, which aims to be a true artificial brain, and Type B, which has a scope limited to performing useful specific tasks intelligently.

Recommendations

Recommendations for enterprises

Type A AI aims to create a thinking machine on a par with human brain capabilities and is still a distant goal not likely to be achieved for many years. Type B AI has a limited but still significant aim to perform specific well-defined tasks in an intelligent way. The progress in Type B is where the excitement exists today, with AI systems available now from startups to well established companies transforming many tasks with rapid automated intelligence.

Every business should consider how AI will affect their domain and should look for opportunities to enhance their field with AI systems. The areas first to be affected will relate to speech recognition, computer vision, and robotics, both physical robots and RPA systems residing digitally. So, for example, enhanced computer-machine interfaces using voice will become available. The 99% accuracy in voice recognition will complete the transformation and is likely to be achieved within the next couple of years. Assisted driving with AI systems will become standard, reducing accident rates. The examples of the startups in this report provide an indication of early breakthroughs and applications.

Businesses, especially large enterprises with sufficient resources, may wish to create their own AI initiatives internally, making use of deep learning frameworks such as Caffe, Theano, Torch, Minerva and others, as well as GPUs and the cuDNN library from Nvidia. Microsoft has made available AI components on Azure for use by developers.

Alternatively, opportunities exist for early adopters of the AI services and products offered by startups and other businesses. There are examples of AI systems to assist in company compliance and regulation, detect fraud, improve security, create advanced human-machine interfaces through speech and vision, assist vehicle drivers, provide assistance to the medical profession, mine big data, and many more applications.

Recommendations for vendors

AI systems pose a threat and opportunity to most business. Because AI has penetrated all types of businesses, at a minimum it will change how we interact with the IT systems. Competitive advantage will go to businesses that can see how AI systems will impact their domain and can exploit the upcoming changes. The market for supporting businesses in working with and building AI systems will also grow. While automation in the workplace will grow and reduce affected jobs, there will be a growth in support jobs, as well as new roles emerging requiring expertise in using AI systems for business advantage.

Deep learning-based neural networks have created a step change improvement in the accuracy of AI systems

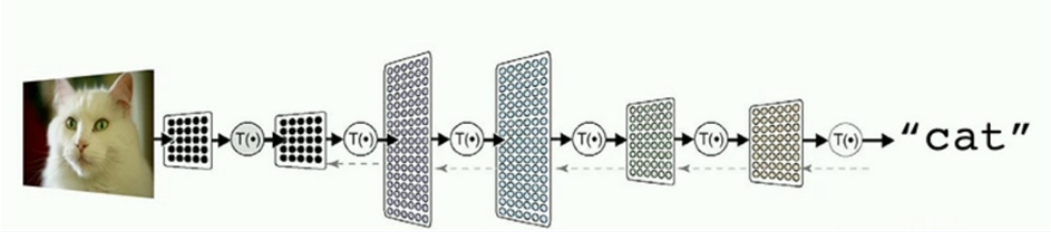
The impact of deep learning

The AI field, and in particular the segment preoccupied with massively distributed parallel systems rather than symbolic computation, has gone through a number of cycles in its history, which started in the mid-1940s. The creation of the backpropagation algorithm in 1986, which allowed quite complex AI models such as neural networks to be trained for real-world problems, led to a new impetus in the field and hype around its capabilities, but once the hype died down the field settled into a plateau and AI techniques found their way into specialized products. One example is the credit card risk profiling system devised by FICO (formerly known as Fair Isaac Corporation), where a trained neural network would detect anomalous spend behavior on a credit card and alert administrators for further investigation. Another example is the LeNet neural network that found applications in check handwriting recognition.

The Backpropagation algorithm allowed a multi-layered neural network (also known as the Perceptron) to be “solved”. In this context solved means trainable, so that the network learns and is able to accurately generalize with fresh inputs not seen during the training phase when the weights in the network are first hardened. These weights are free parameters that start with random values and are gradually fine-tuned during the training to produce a learning capability such as pattern recognition.

In the last five years a new learning algorithm, deep learning, coupled with accelerated running time on Nvidia GPUs has begun to gain momentum and has rekindled a new surge of enthusiasm in neural networks. The learning and generalization capabilities of this new generation of neural networks has raised their capability a notch upward and this is where business-oriented applications are beginning to emerge.

Figure 1: A schematic diagram of a deep learning neural network system



Source: Andrew Ng, Baidu

The most successful deep learning systems learn basic features, such as edges in imaging tasks, and then achieve higher and higher abstractions with progress through the hierarchical system structure (see Figure 1). From our knowledge in neuroscience, this is similar to how the brain processes images. Deep learning face-recognition systems will find a natural recognition of a face which can be achieved without supervision by the human trainer. Figure 2 shows how such a network internalizes an ideal face for which it will trigger the highest recognition.

Figure 2: Deep learning system with unsupervised networks for face recognition

Source: Quoc V Lee, Stanford University and Google

Business applications powered by deep learning are now available

Startups are emerging with deep learning-based applications

The startup field in AI is blossoming, here are some names powered by GPUs and deep learning.

- Chase IT: Intelligent Voice for compliance and eDiscovery solutions
- Clarifai: visual recognition system
- Dato: multiclass classification
- Emotient: emotion detection and sentiment analysis
- Enlitic: medical diagnostics
- Ersaltz Labs: data mining
- EyeEm: photography
- GeekSys: retail store performance management
- HertaSecurity: facial recognition system
- **Ifl**ytek: speech engine
- InsilicoMedicine: genomics and big data analysis
- Jibo: family robot
- iQIYI: online video platform
- Megvii: Face++ platform for face recognition
- Metamind: natural language processing
- Nervana Systems: hardware for deep learning
- Orbeus: image-to-text technology to index videos
- Paracosm: vision for robots and augmented reality
- QM Scientific: shopping intelligence
- Replica Labs: computer vision
- Sensetime: computer vision
- Sogou: Sogou search engine
- Zebra Medical Vision: big data medical imaging

Established companies also working with GPUs and deep learning include Adobe, Alibaba.com, Amazon, Baidu, Cycorp, Facebook, FICO, Flickr, Yahoo!, Google, IBM, Microsoft, Nuance, Scanadu, and Twitter. These players are also buying up startups and others. For example, Google recently acquired Boston Dynamics, an MIT robotics spin-off launched in 1992, and in 2014 it acquired DeepMind, which was founded in 2011.

The auto industry is also investing in these technologies with driver-assisted systems. Tesla Motors has GPUs under the bonnet running advanced software, and Audi, BMW, Volkswagon, and others are also active in this area.

Three examples from the startups listed above will be are examined here in more detail to indicate the type of markets being addressed.

Chase IT

This London-based startup offers Intelligent Voice for monitoring telephone conversations and turning voice into text that is then mined for information. It means that telephone calls can be searched like any text search, and the JumpTo feature provides information on unusual conversations that appear different from the norm in the given environment. The product is in use in London's investment banks where conversations are monitored for compliance reasons.

Clarifai

This service offers image recognition and was the ImageNet winner in 2013, an open competition to test AI systems. The company says its products now go beyond those results, and probably for commercial confidentiality reasons it no longer participates in the ImageNet competition. Based in New York, the startup offers improved speed, vocabulary size, and memory footprint, and has expanded beyond images to extract knowledge from other forms of data. It is possible to see for yourself the service in action by entering an image file or URL and letting the solution return similar images: http://www.clarifai.com/#demo.

Herta Security

Part of the Everis Aerospace, Defense and Security group, the startup originated in Barcelona. It offers fast and accurate video surveillance and access control solutions exploiting deep learning and GPUs. It has international projects that include safe-cities, airports, train and metro stations, prisons, banks, casinos, sports stadiums, shopping malls, military, police and forensic applications. Herta Security has partners in 25 countries.

The following is an example from a mature company.

IBM

IBM Watson, which made headlines as an intelligent machine that won the US TV game show Jeopardy, has progressed from a single machine (a box in a room) to the cloud, where it draws on a collection of AI algorithms and is also scalable on the cloud, with IBM offering commercial cognitive services. It is suitable for a wide variety of applications, and IBM is working with partners to address an increasing number of these, including:

- Medical diagnosis and action: Assists medical professionals diagnosing a particular case and patient to identify a condition and suggest next steps.
- Contact center support: Machine responses offering personalized self-service experience for clients by dynamically developing personal profiles from unstructured data.
- Research and discovery: Identification of rare studies and information sources while building a case for original research, such as, for example, assisting the pharmaceutical industry to discover new drugs.
- Process optimization: Identification of areas for improvement in business processes by analyzing unstructured data that documents and describes process steps and output.
- Fraud and risk management: Identification of early signs of fraud and management of risk in order to lower overall liability and costs of doing business.

The next decade: a market for AI systems worth billions

AI will matter as a result of initiatives in Type B AI that are showing success. These technologies will disrupt existing solutions on the market as AI-based systems outperform conventional technologies. They will be embedded in many applications, services, and devices. These software applications will assist humans and also replace humans in certain tasks that can be automated and require a low level of intelligence but benefit from enhanced (possibly super-human) pattern-recognition capabilities.

Cognitive computing places the emphasis on computation, with advanced computer systems playing a key role. Natural language processing (NLP) is the application of computational linguistics to human language technology, such as, for example, the extraction of structured information from unstructured text. Neural networks are inspired by the brain's neural structure, and run mostly as computer simulations or sometimes on special hardware, and are able to perform advanced AI tasks such as face recognition.

The market for expert AI systems is not new. Early work in the 1990s saw successful expert AI systems, such as medical assistants, but these were rejected by professionals because they were perceived as job threatening. Now the situation has reversed, with professionals swamped with information and finding it difficult to stay abreast of essential new developments in their field. In these situations, an expert AI system can support a professional and make their work successful and their jobs more secure. The use of AI-based face recognition was first used by casinos to identify known Blackjack counters and other inconvenient winners. The technology is now used in surveillance in multiple ways, such as, for example, to recognize when crowds in confined spaces are becoming dangerously dense with a danger of suffocation, or recognizing and tracking individuals.

Technology startups are springing up and also being acquired by the likes of Google, Facebook, and others, introducing a new generation of products that can exploit these AI advances. Many are applicable to data mining, big data, and predictive analytics. In the next decade these technologies will become less expensive and more prevalent in everyday experiences, including home, office, and travel. An under-explored area is the use of AI by artists in the creative arts (computer games and CGI apart), and we may see a new generation of artists inventing novel ways to use AI.

Table 1: 2025 projection: technology analysis matrix for Artificial Intelligence

Geography	Customer	2015-2017	2018-2021	2022-2025
Emerging market	**Business**	**AI-based decision support systems (IBM Watson, for example) as commercial services. Verticals include medicine/healthcare, science research, education, and others.**	**AI advances will feed into robot brains. These machines will begin to be seen in offices and factories and beyond.**	**There will be small improvement increments, still mining the deep learning algorithms. Creative AI will appear where artists explore new possibilities.**
	Consumer	**AI-based consumer products/apps, such as music-recognition apps. AI is already used in Apple Siri/Google Voice.**	**The Intelligent home is one where AI assists and automates, such as to record television programs, to cook, to play games.**	**A breakthrough in instant speech recognition with 99% accuracy is likely, even under noisy environments.**
Developed market	**Business**	**Financial algorithmic trading dominates high-frequency trading, and the**	**Use of advanced AI assistants will be a norm in many products and software**	**Robot brains infused with AI advances will become a normal sight in offices and**

		steady rise in use of AI systems will continue.	**applications, such as data mining.**	**homes.**
	Consumer	**Games have long been an enclave for AI and the ingenuity in AI game opponents will continue.**	**Intelligent personal apps for use on smart devices will be a mushrooming consumer market.**	**The intelligent home will create a market opportunity for AI machines, applications, maintenance requirements, and so on.**

Source: Ovum

Increasing research in the past five years has been in the application of Type B AI to create hardware-based Type A AI-like systems. For example, Neuromorphic chips are being used to emulate as much of the brain's structure as we currently understand. Some initiatives are less "blue-sky" and more commercially focused, such as IBM's TrueNorth chip. The next decade will see such chips assisting AI systems with massively parallel-processing capabilities. Table 1 shows the potential for AI over the next decade.

The combination of deep learning algorithms accelerated on GPUs has been the pivotal breakthrough

The ImageNet competition shows a 26% to 5% error rate improvement in the last five years

The time it takes to train a neural network has an impact on the possible research undertaken. Clearly, the faster the training, the faster is progress achieved in the research. Progress before GPU usage was painfully slow, whereas GPUs now accelerate research. The range of deep learning-based neural network training times and how they affect research is summarized here.

- A month or greater: inhibitive to research, and characterizes pre-GPU days for large-scale networks.
- One to four weeks: High-value experiments only, still a significant lag on research and characterizes pre-GPU days for small to medium size networks.
- One to four days: This is now in the GPU enabled zone. Is acceptable and enables high degree of experimentation.
- Minutes to hours: In the real-time interactive research range, made possible with GPUs. Allows the fastest possible progress.

One technique used in building a deep learning system includes using combinations of four of the major types of neural network learning algorithms: supervised learning, unsupervised learning, reinforcement learning, and recurrent learning. Convolution neural networks (CNNs, which go back to 1995) are a major part of deep learning systems and were found to yield a step change improvement when first used in 2012 by Krizhesvky et al. In 2015, all ImageNet contestants used deep learning with convolution neural networks.

A benchmark in testing AI systems is the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), acompetition that started in 2010. It uses a fresh image library each year designed to stretch the capabilities of the AI systems. Each machine contestant is marked on the top 5 retrievals that are produced by the machine for a test image that it has never seen before. The most recent winning systems are shown in Table 2, and note the score achieved by a high-performing human.

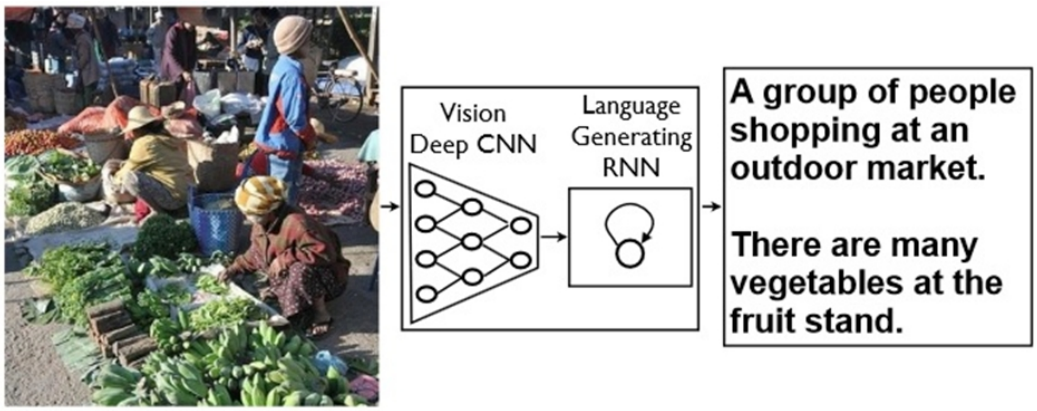
Table 2: ImageNet competition results:

Year	Winning entry	% Top 5 error rate
2010	NEC Labs America	28.2
2011	Xerox Research Centre Europe	25.8
2012	Krizhesvky et al.	16.4
2013	Zeiler/Clarifai	11.7
2014	GoogLeNet	6.7
January 2015	Baidu	5.98
	Human: Andrej Karpathy	5.1
February 2015	Microsoft Research	4.94
February 2015	Google	4.82

Source: Ovum combination of ImageNet results and Jeff Dean slide

Google uses its research in deep learning to power Google Plus Photo search to find images related to a keyword, as well as other applications. For example, Google has a system that observes the real-world environment and reads the name of shops and other signs, and provides a text commentary or translates the text. The most recent examples, from Stanford University and Google, can observe an image, recognize its content, and write a descriptive caption, with a level of accuracy equaling humans (see Figure 3).

Figure 3: Computer vision: Machine image recognition and descriptive captions generated



Source: Oriol Vinyals (Google). RNN is a recurrent neural network.

Nvidia is the main player in the GPU accelerated AI market

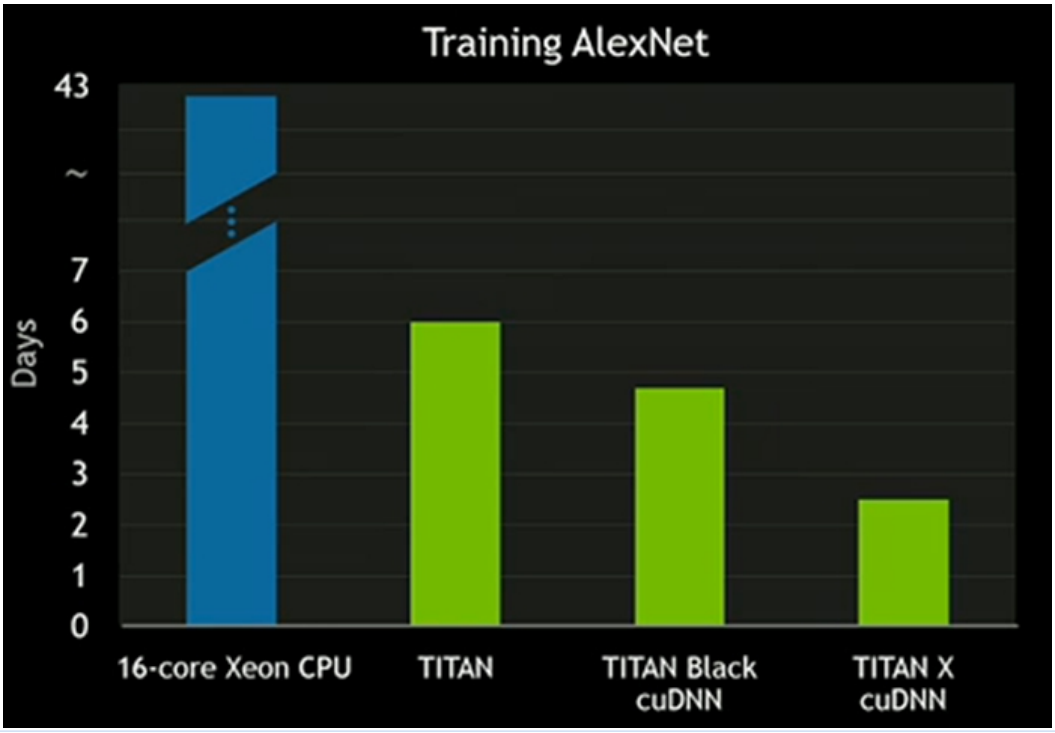
Nvidia has a few years' lead over its nearest rivals in the programmable GPU market with applications in AI, as well as other markets such as high-performance computing. In 2007 it launched its compute unified device architecture (CUDA), a programmable parallel computing platform implemented in GPUs. CUDA programming provides the fastest and most versatile way to access the power of Nvidia

GPUs, with core supported languages C, C++, and Fortran. Third-party wrappers are available for Python, Java, Ruby, Lua, Haskell, R, Matlab, Mathematica, and others.

In 2008 there were 150,000 downloads, 27 CUDA applications, 4,000 academic papers citing CUDA GPUs, 6,000 Tesla GPUs, and 77 supercomputing TeraFLOPS. In March 2015 there were 3 million CUDA downloads, 319 CUDA applications, 60,000 academic papers citing the platform, 450,000 Tesla GPUs in the market, and 54,000 supercomputing TeraFLOPS.

Nvidia announced its latest and fastest GPU at the GPU Technology Conference 2015. Titan X contains 3,072 CUDA cores, 8 billion transistors, 12 GB memory, and can achieve 7 TeraFLOPS in single precision. Its cost is $1,000. When combined with the new CUDA deep learning library (cuDNN), which provides primitives for connecting deep learning frameworks with CUDA, AlexNet, an example of a deep learning system, reduced its training time from a month to 2.5 days (see Figure 4). This is a huge degree of acceleration.

Figure 4: Titan X GPU: training time for the deep learning system AlexNet



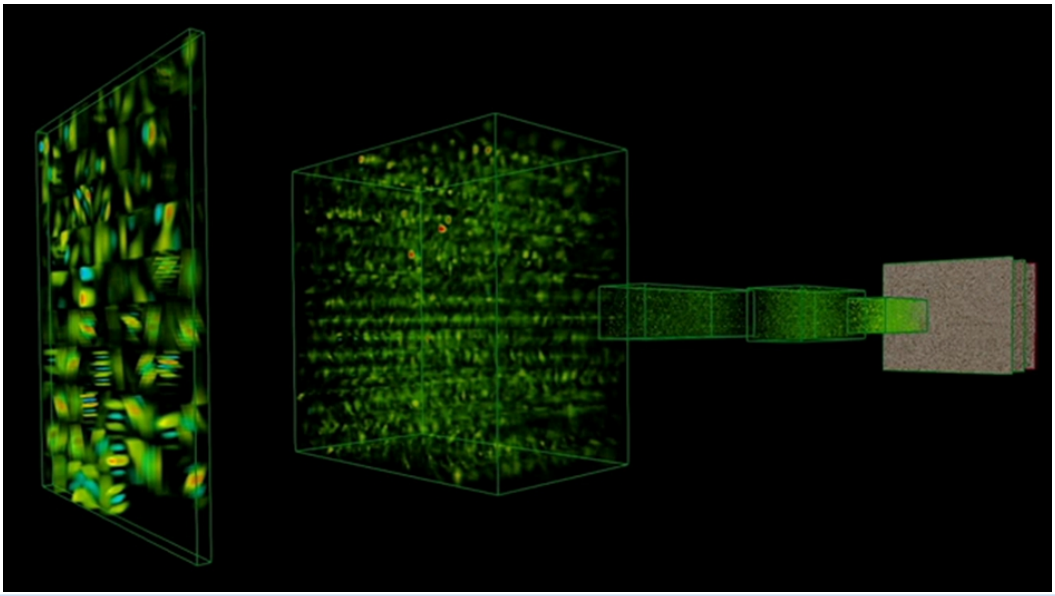
Source: Nvidia

The most popular deep learning frameworks that can be integrated with CUDA via cuDNN include Caffe, Theano, and Torch. The first two use the Python programming language, and the third uses Lua. Figure 5 shows a rendering inside an actual deep learning system of the images being processed at multiple layers, starting with the first layer on the left side which is shown many real-world images at the same time. The connections between the layers are not shown. The final layers on the right with the greatest detail are the first fully connected layers that combine the abstractions produced in earlier layers. This visualization demonstrates the hierarchical processing of images from edges to higher abstract entities.

Nvidia has also launched DIGITS, an open source Interactive deep learning GPU Training System, which is a complete system aimed at data scientists and others, without having to write code. Its features include:

- Visualization of deep neural network (DNN) topology and how training data activates a neural network.
- Management of training of many DNNs in parallel on multi-GPU systems.
- Simple setup and launch.
- Import of a wide variety of image formats and sources.
- Monitoring of network training in real time.
- Open source so DIGITS can be customized and extended as needed.

Figure 5: Visualizing the inner workings of a deep learning system processing images



Source: Nvidia

Open source and sharing of knowledge is helping accelerate AI innovation

Noticeable is the impact of the open source movement on the AI field. All the key academic papers on deep learning and related areas are now openly accessible. Source code is shared between research teams and deep learning open source libraries such as Caffe, Theano, and Torch are helping spread knowledge in AI. This sharing of knowledge is helping accelerate the take-up of these techniques and leading to startups with business applications.

There are two kinds of AI and only one is practical for businesses

Defining AI

AI can generally mean one of two things: Type A or Type B.

- Type A means the creation of an artificial human brain. This is not the simulation of some aspect of a human brain, but instead a fully functioning human-like brain that thinks like we do. It may or may not have consciousness or self-awareness. One can get into philosophical

debates about whether consciousness is necessary for intelligence, so we leave that aside. However, there is one important point: science as it stands today does not fully understand how the human brain works, in fact we do not understand how the most primitive brains function in creatures with a very small number of neurons (worms, flies, for example), let alone in an average human brain with 100 billion neurons. It is unlikely that a type A AI will be built until we at least understand how nature's "wet-ware" brains work.

- Type B means the creation of advanced computer/machine learning systems, pattern recognition systems, and expert systems for largely scientific, medical, and engineering purposes. Type A research typically influences Type B research, but while Type A is still a distant goal, Type B has achieved some notable successes.

Many names used for Type B AI are just synonyms, and reflect the preferences of different research groups including cognitive computing, machine learning, and computational intelligence. The next level down from AI the field is split between symbolic reasoning and massively parallel distributed processing. Much research effort and successes are occurring in the latter branch (Bayesian networks, evolutionary computation, genetic algorithms, and neural networks belong to this group).

Dark side of AI

AI advocates talk of the "singularity" when humans will one day create a genuine Type A breakthrough, then their creation will have the capacity to create/invent the next level of AI brain, and this continues with ever more intelligent AI brains, transcending humans with every step. There is a genuine threat to human existence in such a scenario, if, for example, intelligent machines view humans as competitors to scarce resources. Given that the inner workings of neurons are still a mystery, we do not expect we will witness the AI singularity in the next 20, 30, or more years, but we do believe it is just a question of time and further research until the question of how the human brain thinks is cracked, and then the singularity becomes a genuine possibility. It is likely that in the next decade or two these types of concerns will lead to legislation and control of AI research as Type B successes bring AI to the fore. Advanced AI brain research will require licensing and auditing, and legislation will require in-built safety mechanisms to ensure humans are not harmed. For readers familiar with the "I Robot" science fiction of Isaac Asimov, this is familiar territory. Asimov wrote about future robots that had a set of three in-built laws designed to prevent harm to humans. The paradox here is that it is likely that military robots will be built to replace soldiers on the battlefield.

Appendix

Further Reading

"Nvidia emphasizes deep learning as a future market for GPUs", IT0022-000331, March 2015

"Open source is accelerating artificial intelligence innovation", IT0022-000328, March 2015

Beyond the Hype: Assessing the Evolution of Robotic Process Automation, IT0019-003367, September 2014

Author

Michael Azoff, Principal Analyst, IT Infrastructure Solutions

michael.azoff@ovum.com

Ovum Consulting

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Ovum's consulting team may be able to help you. For more information about Ovum's consulting capabilities, please contact us directly at consulting@ovum.com.

Copyright notice and disclaimer

CONTACT US

www.ovum.com
analystsupport@ovum.com

INTERNATIONAL OFFICES

Beijing
Dubai
Hong Kong
Hyderabad
Johannesburg
London
Melbourne
New York
San Francisco
Sao Paulo
Tokyo