FUTURE OF BIG DATA

Dominique A. HEGER, PhD
President and CEO DHTechnologies
& Data Nubes, Austin, USA

DHTechnologies & Data Nubes, Austin, USA

In this 2015 keynote, Dr. Heger focuses on some of the technologies that have the potential of significantly impacting the roadmap for Big Data. Due to the time constraints imposed onto the presentation, only Deep Learning and Quantum computing are highlighted. It has to be pointed out though that other aspects such as Ambient Intelligence or advanced in-memory computing techniques will have a profound impact on the future of Big Data.

Deep Learning

A typical machine learning (ML) problem scenario is that a system is faced with a large set of data and on its own, is tasked to sort the elements of the data-set into categories. A good analogy would be to ask a child to sort a set of toys without providing any specific instructions. The child may sort the toys by color, by shape, by function, or by something else. The ML scenario described here attempts the same thing just on a much larger scale. A system may be fed millions of handwritten digits and is supposed to guesses which digits look comparable, basically clustering the digits together based on similarity. The essential deep learning novelty is to design and implement models that learn these categories incrementally by trying to identify first, lower-level categories (like letters) prior to attempting to obtain higher-level categories (like words). So in a nutshell, the term deep learning describes a particular (novel) approach to implementing and training artificial neural networks (ANN). ANN's were introduced in the 1950s and so they have been around for a long time. Artificial Intelligence in general has been labeled as an amazingly promising research idea, but the practical deployment did not really happen until a couple of years ago. Simplified, a neural network can be described as a decision-making black box system. ANN's consume an array of values (that may represent pixels, audio waveforms, or strings), execute a series of functions on the array and output 1 or more values as the result. The output normally reflects a prediction of some properties one is trying to estimate based on the input (to follow one of the initial deep learning projects conducted by Google, to determine if an image is a picture of a cat). The functions that execute inside the black box are controlled by (1) the memory of the neural network

and (2) the arrays of numbers that are labeled as the weights that define how the inputs are combined and recombined to generate the results. Processing real-world problems such as Google's cat-detection project requires very complex functions, which implies that the arrays are rather large (sometimes in the range of millions of values/numbers). The biggest obstacle to really using ANN's has been to determine how to set all these gigantic arrays to values that efficiently/effectively transform the input signals into output predictions.

One of the goals and objectives of ANN research is that the networks should be teachable. It is rather trivial (at a small scale) to demonstrate how to feed a series of input examples and expected outputs into the system and to execute a process to transform the weights from initial random values to progressively better numbers that produce more accurate predictions. The main problem is how to do the same thing at a large scale while operating on complex problems (such as speech recognition) where a much larger numbers of weights is involved. In 2012,

Krizhevsky, Sutskever, and Hinton published a paper [ImageNet Classification with Deep Convolutional Neural Networks] outlining different ways of accelerating the learning process, including convolutional networks, smart ways to utilize GPUs, as well as several novel mathematical approaches such as Rectified Linear Units (ReLU) and dropout procedures. The term dropout refers to a technique that avoids overfitting in neural networks. The researchers showed that within a few weeks, they were able to train a very complex network at a level that outperformed conventional approaches used to solve computer vision related tasks. The new techniques outlined by the researches have also been successfully applied to natural language processing and speech recognition related projects and so form the heart of deep learning. With most machine learning projects, the main challenge is in identifying the features in the raw input data set. Deep learning aims at removing that manual step by relying on the training process to discover the most useful patterns across the input examples. It is still required though to design the internal layout of the networks prior to the training phase, but the automatic feature discovery step significantly simplifies the overall process. Deep learning excels at these unsupervised learning tasks but still has much room for improvement. To illustrate, the much hyped Google system that learned to recognize cats outperformed its predecessor by about 70%, but still only recognized less than 1/6th of the objects on which it was trained. Rationally, deep learning is only part of the equation of building intelligent machines. Today's techniques do not allow representing causal relationships, perform logical inferences, or integrate abstract knowledge. The most powerful artificial intelligence systems today use deep learning as 1 element and fuse it (in a rather complex manner) with other techniques ranging from Bayesian inference to deductive reasoning.

Quantum Computing

While the current distributed computing infrastructure (including the Big Data & Hadoop ecosystem) is considered as being very powerful by today's standards, to actually increase the computing power of these systems to address the ever-growing Big Data project requirements, it is necessary to add additional transistors into each cluster node. This effort is getting more and more difficult though as it is already a daunting task to add additional transistors based on the currently available technology. By 2012, the highest transistor count in a commercially available CPU was over 2.5 billion transistors (Intel 10-Core Westmere-EX). Ergo, a paradigm shift is necessary to meet future computing demands. Quantum computing refers to the fusion of quantum physics and computer science and represents a paradigm shift where data is represented by qubits. Unlike conventional systems that contain many small transistors that can either be turned on or off to represent 0 or 1, a quantum bit represents any possible superposition of 0 and 1 with complex numbers serving as the coefficients of the superposition. In other words, A qubit is represented as any point inside or on the surface of a 3D sphere (Bloch Sphere). This property is known as superposition in quantum mechanics (Schrödinger's cat comes to mind) and it provides quantum computers the ability to compute at a speed that is exponentially faster than conventional computers (for some quantum algorithms!). For certain problems such as database search operations or factoring very large numbers (which is the basis for today's encryption techniques), quantum computers can produce an answer many orders of magnitude faster than today's fastest systems (some problems today just cannot be solved within a reasonable time-frame). In quantum computing, a data value held in a qubit has a strong correlation with other qubits even if they are physically separated (Bell theorem - Spooky action at a distance). This phenomenon, known as entanglement, allows scientists to state the value of one qubit just by knowing the state of another qubit. Qubits are highly susceptible to any effects of external noise. Hence, in order to guarantee accuracy in quantum computing, qubits must be linked and placed in an enclosed quantum environment, shielding them from any noise. Currently, quantum computing is still considered an (advanced) research project and no pure quantum computer is commercially available (state 2015). Nevertheless, quantum computing receives significant funding and extensive research is being conducted on the HW as well as the SW (algorithm design) side and substantial advances have been made lately. Google has been using a quantum annealing computing system (from a company called DWave) since 2009 to research highly efficient ways to search for images based on an improved quantum algorithm discovered by researchers at MIT. In 2012, IBM research (based on work by R. Schoelkopf, Yale) derived a qubit that lasted for 1-10,000th of a second. This amazing achievement aids in

lengthening the time-frame for error correction algorithms (to detect and resolve any possible mistakes). Generating any reliable results from quantum computing requires a very low error rate and hence, the IBM achievement is considered a breakthrough.

Author

Dominique Heger has over 29 years of IT experience, focusing on systems modeling, performance evaluation, and capacity planning. Further, he specializes in optimizing Big Data processing environments, predictive analytics, as well as Machine Learning related projects. He is the owner/founder of DHTechnologies (www.dhtusa.com), an IT performance consulting firm as well as Data Nubes (www.datanubes.com), a Big Data and Predictive Analytics company. Bothe companies are headquartered in Texas. He has successfully conducted large-scale IT and Big Data projects for companies such as Boeing, AT&T, LLNL, NERSC, Dell, QLogic, Wells Fargo, EBay, EOG Research, Oceaneering and CERN. Prior to DHT and Data Nubes, Dominique worked for IBM, Hewlett-Packard (at CERN Geneva), and Unisys. Over the years, he has published over 30 papers and books on performance, Cloud, and Big Data related topics with IEEE, CMG, or the IBM Press. Dominique is also the author of the Big Data and Predictive Analytics book (released 2015) that fuses the Machine Learning algorithms, application, Cluster, Cloud, and performance aspects of Big Data projects. He holds an MBA/MIS from Maryville University St. Louis, and a Ph.D. in Information Systems from NSU, Florida. Next to his work in computer science, he very much enjoys spending time with his family, training horses in the Texas Hill Country and team roping.