DARPA Neurocomputing*
Dan Hammerstrom, DARPA/MTO
Arlington, VA USA

Introduction

Intelligent computing has been highly dependent on the Moore's Law¹ bounty. And although there have been algorithm improvements, the 5 orders of magnitude increase in processor speed and memory capacity over the last 40 years account for most of the advances in getting us to where we are today. The DoD is now facing extremely complex and increasingly urgent problems. For example, there are huge quantities of sensor data that require intensive and complex, often real time, analyses. The control of complicated weapon systems has exceeded our ability to write error free software, and cybersecurity threats are evolving faster than we can respond. Addressing these problems requires increasingly more capable intelligent computing. However, we can no longer rely on a stalling Moore's law to get us there. One approach that shows some promise is neural inspired computing, but its near term capabilities are far from obvious. Even so, there are good reasons to be optimistic.

Neurocomputing has been through two periods of significant effort that were then followed by stagnation. We are now experiencing another. The current activity in this area was stimulated primarily by the achievements of "Deep Learning"² with significant breakthroughs in image classification (ImageNet³) as well as other applications such as speech recognition⁴. But there have been other influential neural algorithm approaches as well, including the work of Jeff Hawkins⁵.

In spite of current successes, a key question facing DARPA is, where do we go next? Industry, companies like Google, Facebook and Microsoft, are putting significant research funding into Deep Learning techniques. Since one of DARPA's roles has been to fund longer term, risky, but disruptive research, what then constitutes disruptive research in Machine Learning?

Current algorithms are still challenged in several ways. For example, learning is compute intensive, requiring high precision floating point arithmetic, and it is almost exclusively performed off line. High precision and connectivity requirements still force the use of expensive, power intensive hardware. Important functionality that current algorithms (neural or otherwise) still struggle with is the ability to capture complex structure or relationships in data. Examples include structured data relationships such as one finds in "is a" hierarchies, syntactic relationships and complex rules.

Going forward, many believe that an increasingly important influence towards a solution to these challenges comes from an examination of real neural circuitry. Even though we are a long ways from understanding how the brain works, we now understand enough to allow us to start moving some of the ideas from on-going research in neuroscience into the algorithm space. A merging of existing Deep Learning techniques with biological characteristics, coupled with dedicated silicon technology has the potential for significant disruption.

Biologically Inspired Algorithms

Neuroscience research has led to the exploration of more biologically inspired algorithms for machine learning, adding new features that have not typically been tried in most machine learning algorithms. Examples of such characteristics include low precision computation, one or "one and a half" shot learning, and sparseness in data representations, connectivity and activation. These new "hybrid" algorithms raise the possibility of significantly reduced computation requirements and improved scaling. They promise the possibility of learning on-line and in real time, to allow systems to evolve during operation. They are more naturally suited to capturing temporal features in the data, and, most importantly, there are opportunities for very favorable mappings to specialized silicon structures that promise significant speed and power savings. Recent examples of this include the IBM TrueNorth chip, developed under the DARPA SyNAPSE program, which integrates a million spiking neurons, each with 256 synapses, executing large complex networks in real time using milliwatts of power⁶.

Hawkins and George describe a machine learning approach called Hierarchical Temporal Memory⁵ (HTM) that aims to capture aspects of the biological structural and algorithmic properties of the neocortex. I generalize this concept and use HTM to describe the class of the more biological-like algorithms, to distinguish them from traditional machine learning models. At the time of this writing, DARPA has several studies in progress to explore HTM algorithms and whether they have the potential to meet the DoD requirements mentioned above. These studies are a precursor to a potential new program aimed at evolving existing algorithms, moving them into real applications (a major focus of any neuromorphic program) and creating specialized hardware substrates that promise significant performance increases with ultra-low power requirements for embedded processing.

Furthermore, there do not seem to be many alternate paths to computational intelligence: many existing, non-neural, non-Deep Learning algorithms have significant scaling problems, require extensive manual input, and are falling short of DoD's needs – further motivating the exploration of neural-like approaches.

Finding and Using Structure

As mentioned, perhaps one of the most important desired characteristics, which has received little discussion until now, is the ability to capture the high order structure that most data exhibit, but is typically not captured. An example of this type of work is Zhu, Fathi and Li's research on reasoning about an object's affordance, which goes beyond basic object classification and draws connections between objects and how they are used. Relationships of objects and actions are

* Approved for Public Release, Distribution Unlimited.

represented by a knowledge graph⁷, which can be generalized to sparse graphs. Creating systems that can capture such complex patterns to perform efficient inference remains a difficult problem. HTM models are sparsely activated and connected, and such characteristics lend themselves to naturally describing the sparse, graphical relationships one sees with the high order data structure.

A nice description of how cortical structures build such representations is from the book *Big Brain* by Gary Lynch and Rick Granger⁸, "...brain sensory data is input to random-access encodings in association areas throughout cortex. ... these areas take initial sensory information and construct grammars ... These are not grammars of linguistic elements, they are grammatical organizations (nested, hierarchical, sequences of categories) of percepts – visual, auditory, ... Processing proceeds by incrementally assembling these constructs ... these grammars generate successively larger 'proto-grammatical fragments,' eventually constituting full grammars. ... They thus are not built in the manner of most hand-made grammars; they are statistically assembled, to come to exhibit rule-like behavior, of the kind expected for linguistic grammars."

Hardware

Conventional processors are a poor match to HTM algorithms, since they are constrained in many ways, including the processor/memory partition and limited parallelism, relative to neural models. Likewise they are excessive in key areas, offering more capability than are needed, such as high precision, tiered caches, complex instruction sets, pipelines, etc. DARPA has on-going efforts to build new types of hardware for probabilistic computing. An example is the DARPA UPSIDE program that is demonstrating that analog computing and specialized devices, such as memristors, graphene and spin torque oscillators, can perform inference calculations, which may take the form of a neural network, for image processing applications.^{9,10,11} This is achieved by using the analog properties of the devices to perform the calculation. The inference operations are mapped to devices in ways in which the physics of the devices perform the computation directly. This approach can result in as much as three orders of magnitude improvement in processing speed and an additional four orders of magnitude improvement in power consumption for object classification and tracking applications.

HTM algorithms have digital features that can also greatly benefit from hardware appropriately matched to the algorithm's properties. These properties that would be incorporated into a "Cortical Processor" would include:

1. High connectivity (more extensive than the computational requirements);
2. Local memory and parameter storage;
3. Simple, low-precision computation;
4. Configurable and adaptable; and
5. Sparse activity and sparse local connectivity.

Consequently, in addition to looking at enhancing algorithm capabilities and their impact on a range of applications, there remains a need to also develop HTM specific architectures. Specialized silicon structures can play a key role in a future Cortical Processor program, but will not

necessarily need any specialized devices. We believe that it is possible to create systems with orders of magnitude improvements in speed and power efficiency for machine learning by utilizing standard CMOS technology, without requiring specialized devices, in architectures optimized for HTM execution. An analysis has indicated that roughly six orders of magnitude difference in power efficiency (Giga-Connections / sec / Watt) between a "Cortical Processor" and a desktop CPU (Intel Xeon)¹².

Conclusion

In summary, DARPA is investigating machine learning algorithms and computer architectures that mimic selected characteristics of human intelligence, such as learning and pattern recognition, to address the challenges of data recognition, control and complexity in the continuously evolving environments where DoD systems operate. Learning approaches to date have shown great promise in solving a wider range of problems in less constrained environments, but require high-precision and long compute times, limiting their ability to learn large data sets rapidly or adapt in real time in the field. Neural inspired algorithms allow the use of low precision, hierarchical, temporal memory structures that can rapidly evolve with changing data, minimizing the need for long training times and maximizing rapid, on-line (in the application) real time adaptation. These capabilities coupled with optimized silicon will result in high performance and low power for real-time, embedded system operation for a wide range of applications.

References

- ¹Moore, G., "Cramming More Components onto Integrated Circuits," Electronics Magazine Vol. 38, No. 8 (April 19, 1965).
- ²Bengio, Y., "Learning Deep Architectures for AI," Foundations and Trends in Machine Learning, 2(1), pp.1-127, 2009.
- ³<http://www.image-net.org/>
- ⁴<http://venturebeat.com/2015/05/28/google-says-its-speech-recognition-technology-now-has-only-an-8-word-error-rate/>
- ⁵Hawkins, J., D. George, "*Hierarchical Temporal Memory - Concepts, Theory, and Terminology*", Numenta Inc., 2006-05-1.
- ⁶Merolla, P. A.; et al., (2014), "A million spiking-neuron integrated circuit with a scalable communication network and interface". Science 345 (6197): 668. doi:10.1126/science.1254642.
- ⁷Zhu, Y., A. Fathi, and F. Li, "Reasoning About Object Affordances in a Knowledge Base Representation," in European Conference on Computer Vision, 2014.
- ⁸G. Lynch, Granger, R., *Big Brain*, Palgrave McMillan, 2008.
- ⁹Prezioso, M., et al., "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," Nature, 2015/05/07/print, 521, 7550, p 61-64, 10.1038/nature14441
- ¹⁰Holleman, J., I. Arel, "Low-Power Analog Deep Learning Architecture for Image Processing," in the Government Microcircuit Applications & Critical Technology Conference (GOMAC Tech), St. Louis, MO, March, 2015.
- ¹¹Park, J.L., et al., "Spatially Resolved One-Dimensional Boundary States in Graphene–Hexagonal Boron Nitride Planar Heterostructures," Nature Communications, 2014/11/07/online, 5 (<http://www.nature.com/ncomms/2014/141107/ncomms6403/abs/ncomms6403.html>)
- ¹²Zaveri, M.S., D. Hammerstrom, "Performance/price estimates for cortex-scale hardware: A design space exploration," Neural Networks, (Archival Journal of the International Neural Network Society), Elsevier, April, 2011, Pages: 291-304, DOI:10.1016/j.neunet.2010.12.003.