Sensor Data Processing

Dr. Dan Hammerstrom
Program Manager, MTO

April 2014

Distribution A: Approved for Public Release; Distribution Unlimited
We need to process massive amounts of sensor data of all different kinds
  - Current computational approaches are very limited in processing such information
  - This is particularly true in low power, embedded applications that have limited computational power and memory

Neural inspired architectures are being explored as one potential solution to this problem
  - The potential of neurocomputing remains underappreciated, since there are still so few successful applications that are built primarily from neural components
  - If we are going to leverage neural based computation, we need to better articulate the unique capabilities neural models bring to these applications

In this talk I will discuss the "neural like" DARPA UPSIDE program as one potential solution to the sensor data crisis, I will also “hint” at possible approaches to the “Back End” problem
### PROGRAM OVERVIEW

Exploit non-Boolean computation based on probabilistic inference and emerging non-digital / non-CMOS devices, to enhance the performance and power efficiency of real-time ISR systems and other power constrained applications.

Leveraging emerging devices for ISR processing:

Next Generation Persistent Video Surveillance: Performance requirements significantly exceed current computing capability (Track thousands of moving targets in real-time) mostly due to power constraints.

Probabilistic inference through energy minimization:

Inference module for determining the most likely feature from an energy minimizing matching process. The degree of feature match is an estimate of the probability that the input is a noisy version of the feature associated with closest attractor basin or energy minimum.

### PROGRAM OUTCOMES

**Capability Objectives:**
- Address critical gap between ISR sensor capabilities and the onboard, real-time processing available for image data analysis, such as target recognition and tracking.
- Move beyond the inherent power constraints of scaled digital CMOS (Moore’s Law) to produce a new asymmetry in capability of future weapon systems.
- 1,000X improvement in computing performance, 10,000X increase in power efficiency over traditional solutions.
- Produce systems capable of tracking >10,000 moving targets in real time (current state of art ~100’s).

**Transition:**
- USAF/AFRL, Army Night Vision Lab, MITRE.

### PERFORMERS

**BAE Systems**
- University of Massachusetts
- Johns Hopkins
- UCSD
- Stony Brook University
- SEMATECH

**HRL Laboratories**
- Purdue University
- University of Notre Dame
- University of Pittsburgh
- Intel Corp.
- NIST

**University of Michigan**
- Portland State University
- New Mexico Consortium, Los Alamos Lab

**University of Tennessee**
- Oak Ridge National Laboratory
- Stanford

**Columbia Maryland**
- Amherst, Massachusetts
- Baltimore, Maryland
- Santa Barbara, California
- Long Island, New York
- Albany, New York

**Malibu California**
- West Lafayette, Indiana
- South Bend, Indiana
- Pittsburgh, Pennsylvania
- Santa Clara, California
- Gaithersburg, Maryland

**Ann Arbor Michigan**
- Portland, Oregon
- New Mexico

**Knoxville Tennessee**
- Oak Ridge, Tennessee
Problem: The Computer Efficiency Gap Is Increasing

Data increasing, analysis and compression requirements increasing ...

Digital architectures are not well matched to feature extraction from sensor images
- Images are inherently analog and low precision
- Digital algorithms are created to search for image structures based on existing digital number crunching architectures
- Digital abstractions limit data analysis
  - 100s to 1,000s of digital operations per image activity
  - Wasted energy in excess operations, data movement and precision

Need new computing approaches matched to image processing
- Use the physics of new emerging devices to extract features.
- Data naturally represented in low precision and sparse form are more suitable to devices and efficient for data transfer

Computing directly with devices eliminates multiple layers of hierarchy/inefficiency
Distributed Best Match Association

Steps to Best Match Implementation:

- “Distributed” best match association is implemented by a number of interconnected processing elements.
- Collectively these elements constitute the network state, and an “energy” can be defined and computed.
- By setting the connection strengths appropriately, the training vectors (our kernels) become local minima in the “energy” space.
- All kernels are compared simultaneously in one “energy minimizing” operation.
- The best match is to the kernel that is “closest” given our metric – it is also, under certain conditions, the most likely probabilistically (which is “inference”).
- This is called an “attractor” memory model.

\[ E = -\frac{1}{2} \sum w(i, j)y(i)y(j) \]

Constraints are “soft” and can be represented as an energy “field” which the network tries to minimize.
Key Concept: Association Via Emerging Devices:
The Hardware is the Algorithm

Device Physics

- Coupled devices operating at very low power levels

Interacting Nonlinear Systems

- Spontaneous synchronization corresponds to a stable minimum in an energy space
- The coupling coefficients create a controlled energy surface with specific local minima that models a probability distribution

Bayesian Inference

- The system converges to the “closest” energy minimum, which is generally the most likely vector in a Bayesian sense

Nano-device Array

Associative memories approximate probabilistic inference, Giga-ops Performance and micro-watts of power

Distribution A: Approved for Public Release; Distribution Unlimited
New Paradigm – Non-Boolean, Probabilistic Computing

1. Computing occurs by the physics of the devices (highly parallel)
2. Devices perform the computational equivalent of hundreds of discrete digital operations
3. The model can be configured into hierarchies that accomplish most of the computational work required by the application

Example: Find Features in Sensor Data (7x7 Gabor Edge Finding, 10 Giga-pixel Array)

**Boolean Computation**
- Processor: Intel 6 Core i7, GOPS: 6.7
- 1 inference is 140 operations/kernel, 24 kernels are compared / pixel
- GOPs/watt: 0.1
- Compute time = 7,700 sec
- **460 kilo-joules** (60 watts for 7700 seconds)

**Analog Direct Device Computation**
- Processor: 10 X 10 Array of coupled oscillators Giga-Inferences/sec = 400 (56k GOPS equivalent)
- Compute time = 0.04 sec
- **430 milli-joules**
**UPSIDE: Performance Goals**

**UPSIDE Goals:** 3 orders of magnitude in throughput, 4 orders of magnitude in power efficiency, no loss in accuracy

**DoD Sensing Requirements (Notional)**

~100 Ops/nsec consuming µwatts

**UPSIDE: Emerging Device**

**UPSIDE: Mixed Signal CMOS**

**COTS Efficiency Ceiling**

**Processor Capability**

**Year**

**UPSIDE Program Tasks**

**Image Processing Pipeline**

*An application driver*

- Recreate the traditional image processing pipeline (IPP) hierarchies of Inference Modules (IM)

---

**UPSIDE Task 1**

**IM Development & IPP Implementation**

- Recreate the traditional image processing pipeline (IPP) hierarchies of Inference Modules (IM)

---

**UPSIDE Task 2**

**Mixed signal CMOS implementation of computational model and system test bed**

- Design and Fabricate a mixed signal representation of the Inference Module in state of the art MS CMOS
- Implement a test bed system using MS CMOS
- Validate against IPP simulation
- Simulate the mapping of the Inference Module to specialized devices
- Determine systems level performance-price
- Show simple circuit operation

---

**UPSIDE Task 3**

**Image processing demonstration combining next-generation devices with new computation model**

- CoFe2O4 epitaxial nanopillars fabricated on MgO, R. Comes, J. Lu, and S. Wolf, University of Virginia, unpublished

---

*CoFe2O4 epitaxial nanopillars fabricated on MgO, R. Comes, J. Lu, and S. Wolf, University of Virginia, unpublished*
Data become more “knowledge / context” intensive, containing both spatial and temporal information, as they move through the pipeline.

The “structure” of the data (high order correlations) carries the information.

Current computational approaches do not adequately represent complex spatial and temporal data, limiting the ability to effectively perform complex recognition for important DoD tasks such as anomaly detection and scenario prediction.

And we also need rapid learning.
Can We Better Leverage Biological Computational Principles?

Six principles have been proposed as fundamental to biological and machine intelligence.

1) Sparse Distributed Representations
2) Sequence memory (spatial/temporal patterns)
3) Hierarchy of memory regions
4) On-line learning from streaming data
5) All regions are sensory and motor
6) Attention

Engineer theoretical operating principles of neocortex in hardware.

Key Ingredients to “Cortical” Algorithm, HTM

- **Sparse Distributed Representation**
  - Associative memory model for best match pattern retrieval improving computational efficiency and scalability

- **Spatial/Temporal Sequence Memory**
  - Learns complex sequences of data predicting future inputs and detecting anomalies

- **Hierarchical Structure**
  - Builds rich models with many layers of abstraction ("patterns of patterns")
  - Feed forward, feedback

- **On-Line Learning**
  - Adaptation of network based on acquired data
**Object Detection**: Capture complex structure to allow recognizing parts and their relationships.

Higher Complexity allows for differentiation of more object classes even when very similar:

- “A Bike” vs “Klein Quantum Race”
- “A Car” vs “White 2013 Toyota Tundra”
- “A Person” vs “Joe Smith”
• “... the front end circuits of the brain ... specialize in their own particular visual and auditory inputs, the rest of the brain converts these 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
Identifying Human Actions and Scenes

- Using Hollywood 2 DataBase
  Dataset with 12 classes of human actions and 1466 video clips.

- Present sets of sequences (edge-filtered, binarized) to HTM model

- Characterize/analyze features learned

- Most current Machine Learning technique do not handle temporal data very well

- When operational performance comparisons will have to be made with traditional video analysis techniques

*Seedling performed by Neurithmic Systems*
Mapping to Hardware

HTM requires hardware matched to the algorithm’s needs
1. High Connectivity (more expensive than the computational requirements)
2. Local Memory and Parameter Storage
3. Simple, Low Precision Computation
4. Configurable / Adaptable
5. Sparse activity

Conventional processors are a poor match:
• Constrained: processor/memory partition, limited parallelism
• Excessive: high precision, tiered caches, complex instruction sets, pipelines, etc.

Custom architectures can address HTM requirements:
• High-risk exotic devices unnecessary
• Utilize conventional CMOS fabrication optimized for HTM architecture/computational model
• Can benefit from latest advances in CMOS.
**Hardware Comparisons**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Platform power</td>
<td>490 watts</td>
<td>0.21 watts</td>
</tr>
<tr>
<td>Training Time</td>
<td>6 days</td>
<td>20 Minutes</td>
</tr>
<tr>
<td>Images / second</td>
<td>2.3</td>
<td>1000</td>
</tr>
<tr>
<td>Joules / image</td>
<td>212</td>
<td>1,000,000x less power</td>
</tr>
</tbody>
</table>


**Custom Digital Processor**
- Assumed at 14nm, SRAM per processor node (PN)
- On-chip routing, realistic cortical model

**Modular “Tiled” Architecture**
- 400 PNs per chip
- 600 nodes / PN
- 0.21 Watt/chip
Large-scale DoD application

*Emphasis on data analysis and adaptation*

**Wide Area Motion Imagery (WAMI)**
- Scenario Awareness and Prediction

**Intelligent Signal Processing and Data Sharing**
- Scalable Architecture: Soldier ➔ Large systems
- Sensor Data Fusion and System Collaboration

**APPLICATION DRIVEN**

*ARGUS-IS*

**Scenarios Awareness**

*Emphasis on adaptive control and sensor data processing and reduced dependency on hand written code*

**Autonomous Vehicles**
- Autonomous decision making
- Motion control
- Task Adaptation

*Embedded On-Platform Computation*

*LOCKHEED SMSS UAV*

*North West Aerospace Alliance UAV - UK*

*ONR Large Displacement Unmanned Underwater Vehicle (LDUUV)*

Distribution A: Approved for Public Release; Distribution Unlimited
Example DoD Application – ARGUS IS

Need to mature algorithm(s) and develop scalable hardware to address realistic DoD applications.

DoD application example: ARGUS (WAMI, 1.8 Gpixel sensor) seeks to track 100,000 moving objects in a next-generation wide area imaging (WAMI) system (currently, with analyst assistance they are tracking ~100 object simultaneously)

• Requirement is to identify $O(10^6)$ complete tracks per day (large urban area)
• $O(10^6)$ tracks requires $O(10^4)$ analysts to process a single day’s data
• Tracks are identified by combined spatial / temporal features and are a common example of complex objects
• UPSIDE will increase the load on the back-end

ARGUS data with target/track overlay

Need to automate the pipeline, reducing burden on human analysts.
Objective: Push the limits of algorithm scaling to introduce machine learning to the most challenging DoD sensor processing problems.

Intelligent Signal Processing and Data Sharing

Data Fusion from Different Sensors

- Not possible with any neural algorithms today, nor with traditional techniques
- Creates the capability of using learning systems to model and control complex systems.
- Helps manage signal and system complexity by automating higher order relationships.

Wide Area Motion Imagery

- Surveillance Imaging
- Scenario Awareness
- Tracking convoy of vehicles
• Conceptually one can think of “computational intelligence” as a spectrum
• Among other things we need to do “small c” cognition before we move on to “Big C” cognition
• And, while cognition remains our goal, I believe that it is possible to build very useful systems with the technology we have now
• We have to build this field one brick at a time, moving incrementally to the right

Computing Is currently here

Small “c” cognition: most mammals

Big “C” Cognition: humans

Our Goal

A Rock

A long way!

A long way!

The Krell

Increasing Intelligence

Distribution A: Approved for Public Release; Distribution Unlimited