Abstract

Computing hardware will have sufficient computational ability and memory to match human performance around 2020-25, as measured by CPU capability and memory. The fields of machine vision, computational audition, robotics, and computer science are separately developing components of intelligent systems that, when combined, may begin to bear resemblance to human intelligence and sensory capabilities. Building robust and intelligent sensory systems for multi-source environments requires us to go beyond traditional engineering assumptions about resolution, data reduction, and scope of search spaces, in ways that are foreshadowed by study of biological systems – these are forces that push in the direction of systems with very high compute loads.

Implementing such compute-intensive systems with low power and low latency for commercial applications may require us to develop new architectures (or extensions of existing architectures) with pipelined and/or dedicated structures to perform important compute-intensive functions more efficiently in parallel that would otherwise have to be performed serially in a conventional architecture. Efficiency improvements of this sort trade off against flexibility, in the presence of return-on-investment and time-to-market constraints in a commercial environment, and there will be a unique set of successful design principles that will emerge as we build systems on this new scale. Those design architectural principles will exploit low-level device capabilities for high-level system benefit.
Overview

- Approaches to building robust, intelligent sensory systems
  - Computational models based on engineering intuition
  - Reverse-engineering based on neuroscience
- Compute / Memory / Power / Funding Requirements
  - When will we be able to solve commercially significant problems?
- “Single-source Recognition” is only a part of the sensory problem
  - A large part of the brain is devoted to scene analysis / source separation
  - Resolution requirements are different for transmission, separation, and recognition
- Demonstrations from real-time audition and vision
- Architectural Implications

Approaches to building robust, intelligent, sensory systems

- Common Long-Term Goal: to build a machine with human-like intelligence, capable of functioning in the real world
- Many valid approaches for a goal so complex
- Computational Models based on Engineering Intuition
  - Neural networks, machine learning, genetic algorithms, fuzzy logic, rule-based AI, computational audition/vision/neuroscience.
- Reverse Engineering based on Neuroscience
  - Construction of biologically verifiable, real-time, high-resolution representations, advancing from the sensors inward, in active, long-term collaboration with neuroscientists.
  - Real-time, high-resolution, systems-level computational neuroscience, in a sustained program to actively build a working system to solve real problems.
- In practice, both will contribute and be useful
What can Neuroscience tell us about building robust, intelligent, sensory systems?

- $10^{11}$ neurons, $10^{14}$ synapses
- $10^{16}$ Ops/s, $10^{14}$ MByte
- $20$ W power consumption
- $10^6$ GOPs/W efficiency (compared with $3$ GOPs/W for current HW)
- $V_{dd}(\text{brain})=80$ mV accounts for nearly 3 orders of magnitude of power efficiency, trades power consumption against area/cost/yield $P=CV^2f$
- Vast variety of cell types
- Thousands of modules/regions being studied by

- $30,000$ neuroscientists
- No individual knows everything.

Dramatic increase in knowledge in last 20 years...

(van Essen and Anderson, 1990)
(Douglas and Martin, 1998), (Calvin, 1996)
Collectively, do we know enough about the brain to begin building a realistic model?

When will computers be capable enough?

(Ray Kurzweil, The Age of Spiritual Machines, 1999)
When will computers be capable enough?

Are there fundable, commercially viable applications for the major pathways (audition, vision, motor)?

Commercial Example:
Noise Robustness for Speech Recognition and Telecom
Resolution Requirements are different for Multi-Source Separation than Single-Source Recognition

“The selection of the best parametric representation of acoustic data is an important task in the design of any speech recognition system. The usual objectives in selecting a representation are to compress the speech data by eliminating information not pertinent to the phonetic analysis of the data and to enhance those aspects of the signal that contribute significantly to the detection of phonetic differences... Compact storage of the information is an important practical consideration.”
-- Paul Mermelstein (inventor of mel-frequency cepstral coefficient), 1980.

“The very purpose of phonetic classification in ASR is a significant reduction of the information carried by the speech signal. Thus, the front end processing should be supportive of this task.”

A Visual Analogy

- Voiced speech contains fine structure due to pitch, speaker ID…
Voiced speech contains fine structure due to pitch, speaker ID…
Blurring eliminates fine structure, making pattern matching easier…
And justifies downsampling, which reduces downstream system cost.
Here is a list of states: Alaska, Ohio, Florida, Nebraska, Delaware, Kansas, Wisconsin.
A Visual Analogy

Low-Res OK for single-source

Here is a list of states: Alaska, Ohio, Florida, Nebraska, Delaware, Kansas, Wisconsin.

Hi-Res Needed for multi-source

Here is a list of states: Alaska, Ohio, Florida, Nebraska, Delaware, Kansas, Wisconsin.

Information Flow for Single-Source

Message 50 b/s

Speech signal 40 kbps

(Hermansky, 1998)
Information Flow for Multi-Source

Architectural Implications

- Cortical functions: extensive pattern match, hypothesis generation and pruning, object tracking, HMM/Viterbi search, associative memory
- High-res feature detection, cross- and auto-correlation, and post-processing
- High-resolution sensory pre-processing
Architectural Implications

Human-Level Performance on Real-World Multi-Input Sensory Processing will require:

- Further algorithm development to define robust computational pipeline (analogous to graphics rendering pipeline)
- Real-time Hearing will likely require 10-100 GOps range
- Will need efficiencies better than 3 Gops/W for commercial acceptance in many applications
- Priority on fast-turnaround design, high-res visualization, real-time operation, validation on large data sets for robustness, low latency.
- Power consumption, cost reduction are later-stage optimizations once the key operational principles are understood
- Will favor parallel, pipelined, low-clock-rate, low-voltage hardware-accelerated architecture to bring power consumption to reasonable level, but power consumption will trade against chip area and flexibility
- Best implementation model may be GPU – parallel compute for dedicated processing pipeline