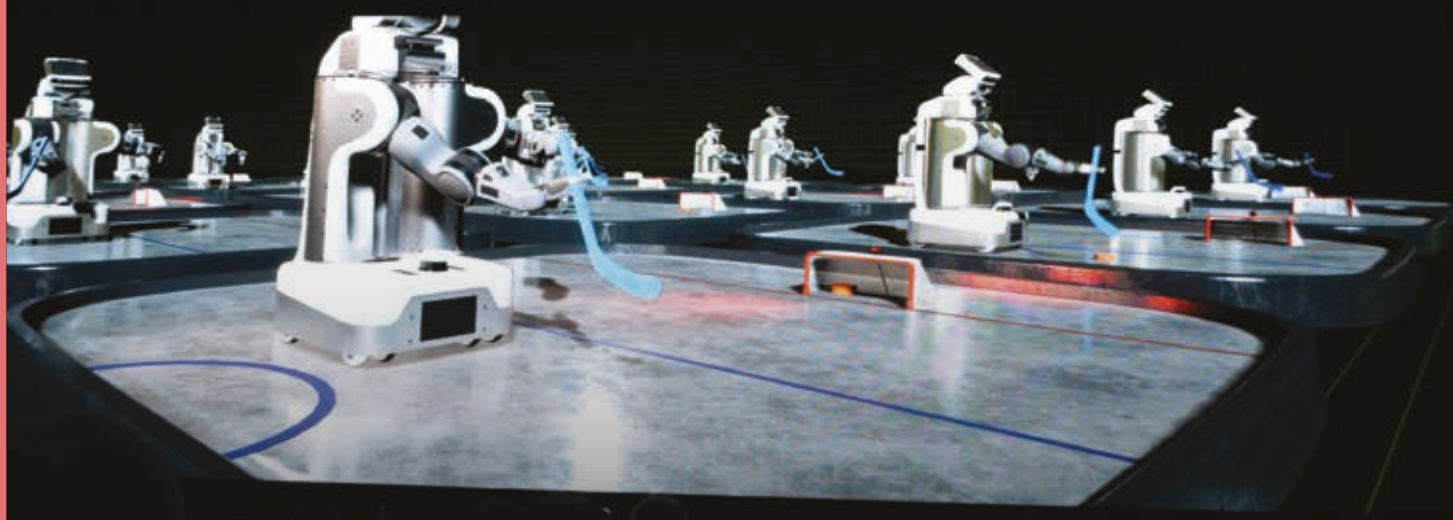

Figure 1: Simulation-based training allows robots and other AI-based applications to build deep, complex, knowledge bases that include worst-case conditions before they are deployed in the real world. Image source: NVIDIA

How GPUs, AI and Deep Learning Make Smart Products Smarter

By Alianna J. Maren, Ph.D.

AI vs. Standard Computing

Advances in artificial intelligence (AI), advanced computing architectures, and low-cost, high-performance sensors are enabling the development of a growing number of commercial applications for autonomous vehicles, drones, robots and other so-called "autonomous edge devices". Since AI combined with deep learning (DL) is a relatively new field, we need to first understand how AI/DL systems differ from conventional embedded systems.

Jensen Huang, co-founder & CEO of NVIDIA, summarizes the difference between AI and standard computing with this phrase: "AI is just the modern way of doing software." Since AI-based systems are, at their core, still digital computers, the tools used to develop AI/DL applications have much in common with those used to program conventional computing applications. There are still functions and many aspects that programmers will recognize.

However, there some aspects of how AIs work, and how they are prepared to support a specific application, that will seem strange, if not downright alien, to the average code-slinger. These differences arise from some basic characteristics of an AI:

(1) A world model that provides the AI with a "machine's-eye view" of the world it is operating in.

(2) The AI's ability to learn from new information and update its world model.

(3) The AI's ability to make inferences that allow it to deal with situations that it has not specifically observed or learned about before.

These unique characteristics are responsible for one of the most fundamental differences between creating traditional software and AI development. When programming a conventional computer, every response has to be pre-specified by the programmer. In contrast, a good deal of programming an AI involves allowing it to "learn" from examples of situations it will encounter, and training it to produce the responses it is expected to give (Figure 1).

The ability to learn and adapt without additional programming enables AI-based systems to do many things that were difficult or impossible to do with conventional computing. It will also become apparent that AI systems require much more powerful computing elements and present the developer with a number of new and complex technical challenges.

A Machine's-Eye View

With AI, there are aspects where the program will learn what to do based on training examples set up by its developer. Part of the developer's job is to assemble sufficient instances of the diverse kinds of data that the program can expect to find, and identify what the desired results should be for these different cases. The AI program learns to perform correctly through using one of several different possible learning algorithms.

An AI's understanding of its environment relies on its world model, a complex data structure that that contains its knowledge about things and relationships between things. For example, an autonomous vehicle's world model will know a great deal about roads, signs, other cars, and everything that it can reasonable expect to encounter on the road. Its world model also contains information about relationships between things, such as the distance between itself and other cars, as well as how that distance is changing over time.

An AI builds its world model from a combination of information generated by the developer and raw sensor data that it collects, correlates and assembles into a coherent set of things and relationships.

In order to function effectively, the AI must use the data it collects and analysis of its past performance to consistently update and refine its world model. In real world/real-time applications, such as autonomous vehicles, the AI updates its model based on inputs from multiple sensors, many of which provide visual data from cameras, or image-like inputs (e.g. LIDAR, RADAR).

This task of assembling a unified 'picture" from a heterogeneous collection of sensors is known as sensor fusion, a technique that will be discussed in greater depth later in this article.

The Power of Inference

The types of AI/DL applications currently under development will not only be required to learn responses to a wide range of known conditions, but to also use inferences to synthesize new responses to effectively deal with novel situations it's never encountered. This involves applying a method (or multiple methods) for identifying the situations that it has already learned which are most relevant to the new situation, and then interpolating between them.

Making reliable inferences is especially challenging because it requires the AI to identify the situations it already knows about that are the closest match to its current inputs using information that noisy or incomplete. In the case of an autonomous vehicle, for example, a sudden change of pixel values between two objects that it has previously identified as being parked cars can have multiple interpretations and correspondingly different responses. If it infers that the newly-detected features are a paper bag blown by the wind, it will have a very different response than if it is a child.

The algorithms used to extract inferences from a model typically require a great deal of computation. Some of these tasks may exceed the processing abilities of the AI's on-board resources and must be off-loaded to more a powerful system

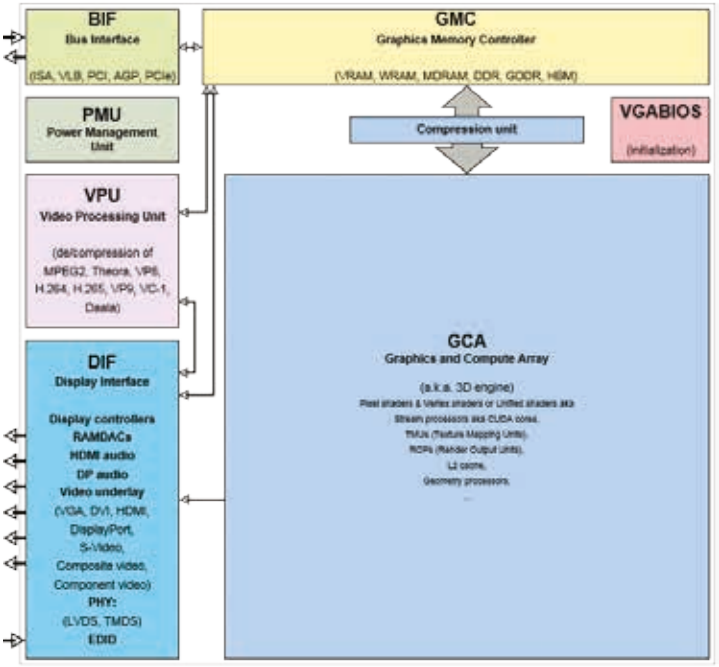

Figure 2. A block diagram of an early GPU (Radeon 9700). Image source: ScotXW, courtesy of Wikipedia

(typically cloud-based) via a cellular connection or other wireless link.

Developing strategies for efficiently segmenting computing tasks between local and remote resources present several challenges addressed later in the article.

GPUs – Not your Grandma's CPUs

Artificial Intelligence's evolution from an academic curiosity to a commercially viable technology has been largely due to the past decade's advances in computing hardware. Arguably, the most common AI-friendly computing architecture is the GPU (graphics processing unit), originally developed as an array processor, optimized for the pixel and vector manipulation tasks associated with graphics and video acceleration (Figure 2). Those same capabilities are also essential for accelerating neural networks, which are essentially vector computations.

NVIDIA, in particular, has spearheaded the GPU's evolution with a series of processors, each tailored to different processing environments and purposes. Among these are Xavier, a system-on-a-chip (SoC) designed for autonomous cars, which will be capable of running at 20 TOPS (trillion operations per second), while consuming only 20 watts of power. The Xavier (Figure 3) integrates the new NVIDIA GPU architecture called Volta that integrates a custom 8-core CPU, as well as a new computer vision accelerator.

NVIDIA also pioneered the use of software tools like CUDA, a parallel computing platform and application programming interface (API) model that allow developers to create powerful GPU-based applications without having to master the intricacies of the GPU's unique architecture and command set.

Autonomous Vehicles Drive Sensor Fusion

The autonomous ground-based and aerial vehicles, expected to dominate transportation by 2030, continue to be the earliest and largest markets for GPUs and AI technologies. All major

automobile manufacturers are racing to incorporate AI and deep learning (DL) into their cars, with plans to deliver vehicles with SAE (Society of Automotive Engineers) Level 3 (partial automation) and potentially Level 4 (conditional automation) capabilities by 2020.

Although rapid advances in AI and DL algorithms are spearheading the transition to autonomous vehicles (Figure 5), this transformation would not be possible without the evolution of sensor fusion, which is largely taking place within the vehicle itself. Until now, the conventional embedded systems used in vehicle control applications processed sensor data on a distributed web of microprocessors, each associated with one, or a handful of sensors. In contrast, sensor fusion brings the raw data from the 60 – 100 sensors found on a typical car onto a single processing platform.

Tomorrow's fully-autonomous vehicles are expected to employ 2X-4X more sensors, needed to support advanced functions like include ultra-precise vehicle location and complete awareness of its surrounding environment.

Making sense of the flood of data arriving at widely different rates and latencies requires the use of an onboard sensor fusion platform to perform a series of challenging tasks, beginning with co-registration of raw sensor data, low-level feature detection (edges and blobs), and identifying preliminary feature correspondences. The platform then associates the edge and blob features, and fuses them to create preliminary objects that are then analyzed by a succession of image understanding algorithms.

At the lower processing levels, edge and feature extraction algorithms, such as Convolution Neural Networks (CNNs) are among the most useful methods (Figure 5). The higher-level processes, in particular, require the use of inference, an AI method explained earlier.

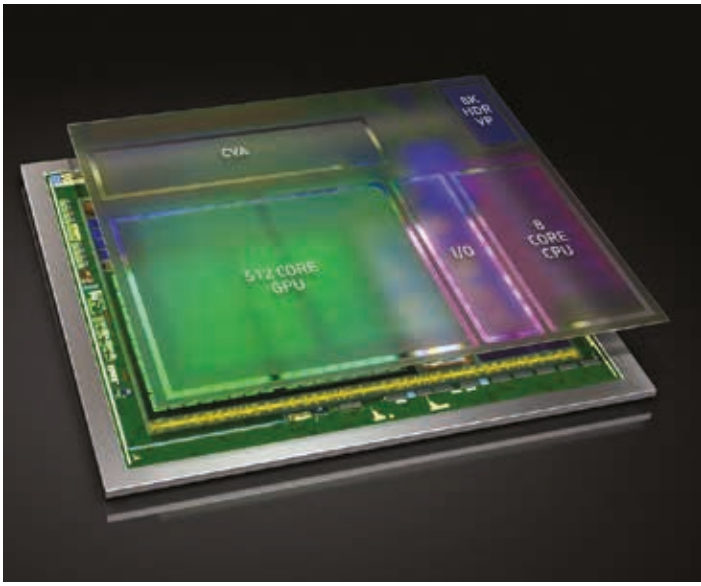

Figure 3. NVIDIA's Xavier System-on-Chip. Image source: NVIDIA

Segmentation Strategies

Autonomous vehicle cost and performance can be dramatically affected by decisions about what processing tasks can (or must) be done onboard the vehicle itself, and what can be performed on the cloud, or pre-trained in a

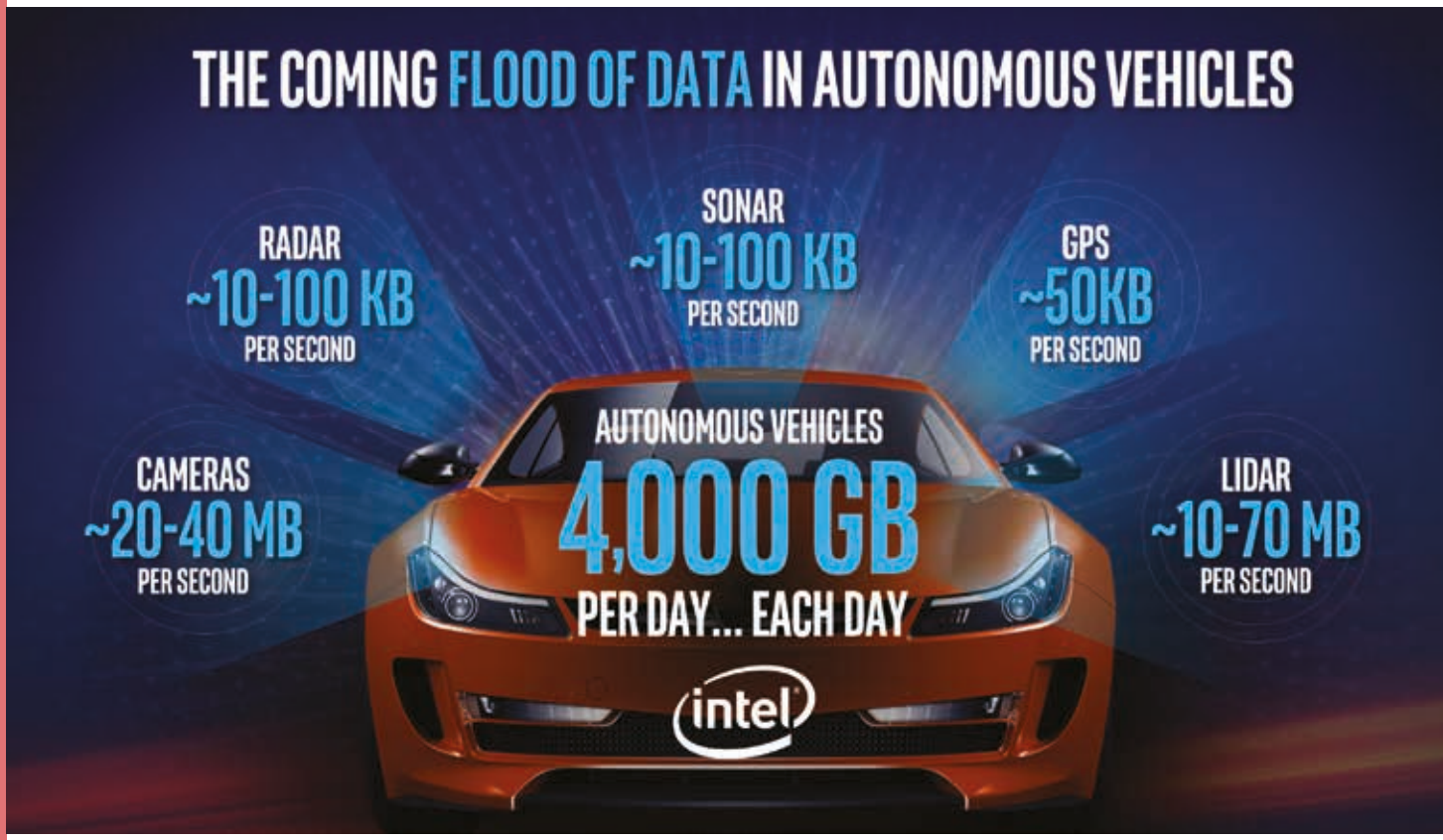

Figure 4. An autonomous vehicle will use a combination of sensor fusion, artificial intelligence, and deep learning to make sense of the flood of data produced by its sensors. Source: Intel

datacenter. Segmentation is often applied to inferential tasks because of their high computational requirements. Time-critical inference tasks should be performed on the mobile device, whenever they don't exceed its processing or storage capacity. Other inferences that are less time-critical, or require a larger knowledge model may be shipped off to the cloud.

Likewise, some of the sensor fusion tasks in cars, drones, robots and other autonomously-mobile products can be done remotely, or efficiently handled by a pre-learned knowledge base, but most sensor fusion must be done onboard and in real time in a centralized hub on the vehicle or robot itself.

NVIDIA has addressed the issue of segmentation with a multi-pronged approach, developing both an onboard processor for AI-based sensor fusion and autonomous guidance and control, and an architecture (together with Microsoft) for a cloud-based AI data center, where the deployable GPUs can be trained, and where the extensive knowledge model can reside. The onboard processing system is based on the Xavier SoC mentioned earlier in this story. Billed as "the world's 1st AI car superchip," Xavier will be trained offline in a datacenter.

AI-Capable GPUs Enable Sensor Fusion

Since nearly all AI-enabled edge applications require some type of sensor fusion technology, many vendors are scrambling to deliver products and platforms to meet the demand. Three of the world's leading chip makers (Intel, Qualcomm, and NVIDIA) and specialty-focused Mentor Graphics Corp. have recently released integrated sensor fusion platforms. These platforms differ from previous sensor-processing systems in two major ways: (1) they accept raw data from multiple sensors, and (2) use neural networks (often deep learning) and other AI algorithms to process and integrate it into a useable form.

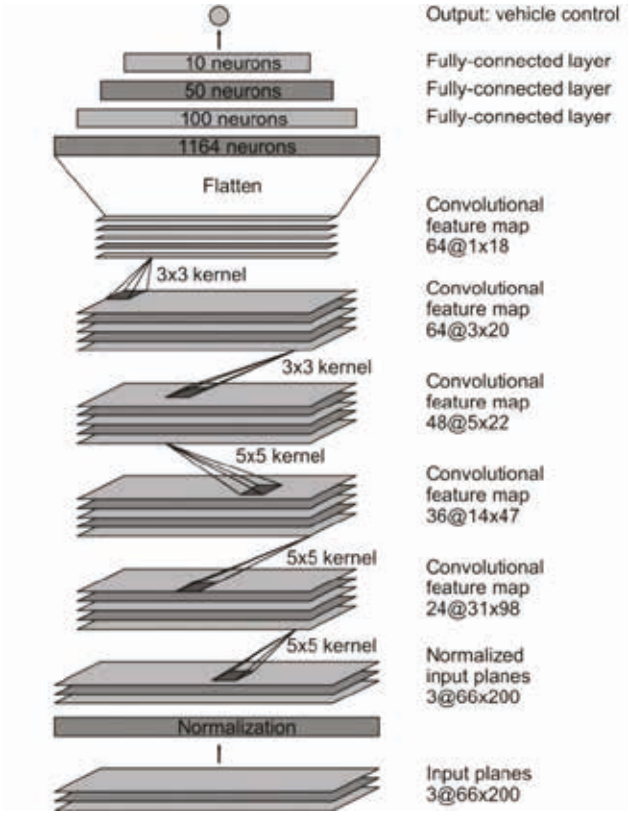

Figure 5. A Convolutional Neural Network architecture. The network has about 27 million connections and 250 thousand parameters. Source: NVIDIA

Qualcomm's Drive Data Platform is designed for easy insertion into the autonomous vehicle supply chain. Based on the Snapdragon 820, it makes extensive use of the Snapdragon Neural Processing Engine (SNPE) to process, fuse, and interpret multiple streams of imaging data generated by camera, radar, and LIDAR.

Meanwhile, Intel has made its own bid to gain traction in the vehicular sensor fusion platform market with the acquisition of Israeli-based Mobileye. Their 5th-generation System-on-a-Chip (SoC) is expected to be deployed in autonomous vehicles by 2020.

NVIDIA's Drive PX2 platform is the product of a collaboration with Bosch, and is based on its forthcoming Xavier device.

AI for Home Security and Appliances

One of other large markets anticipated for AI technology is in home security and home appliance devices. For these household-level edge devices, sensor fusion will be as important as it is in autonomous vehicles. It will also introduce some interesting new types of sensor modalities. For example, Audio Analytic has created a technology that can discern sounds that have "non-communicable intent" from both ambient noise as well as intentional sounds, such as speech and music.

This technology enables a sensor fusion AI system equipped with microphones to monitor a household, office or industrial building and identify any sounds that require an active response, such as crashing glass inside a home, or a siren heard outside a car, even when they are faint, or nearly drowned out by background noise. The system's response could range from collecting more data to sounding an alert to driving a car to the side of the road and allowing an emergency vehicle to pass.

Robotics and Smart Manufacturing

AI and its associated technologies will also accelerate the evolution of the industrial robots that are already making inroads into the factory labor pool. One example is NVIDIA's Isaac robot simulator that changes how robots learn to perform complex tasks (Figure 1). The AI-based software platform lets development teams train and test robots in highly realistic virtual environments. Once trained in the virtual environment, its knowledge can be imported to other models and variants of the original design.

AI Everywhere?

"AI everywhere" is becoming a common phrase as AI finds its way into applications as diverse as personal assistants, personalized recommendation apps in online vendors, home security, robotics, and smart manufacturing. Of all the technology evolutions over the past decade, AI is the one most likely to impact the greatest range of products and their design.

Reader Note: An extended version of this story, complete with longer explanations, more graphics, and extensive references, is available on the PD&D web site at **http://tinyurl.com/PD-D-1709AIStory.** PDD