A Machine Learning Application Landscape

And appropriate hardware alternatives

Summary

2016 was a strong year for Machine Learning (ML) and Artificial Intelligence (AI) with many high tech firms claiming that they are now an "AI Company", notably Amazon, Baidu, Facebook, Google, IBM, Intel, Microsoft, NVIDIA, and Tesla. In 2017, the field will broaden to include AMD, Qualcomm, and Xilinx. Moor Insights & Strategy (MI&S) expects Machine Learning based AIs to begin to expand from cloud-based services and applications (primarily internal to the web company's operations) to specialized applications and edge devices, providing intelligent solutions for a wide range of enterprise, consumer, and industry-specific applications. This MI&S analysis segments these emerging AI applications and explores the underlying hardware required to run these cloud, edge, and hybrid applications.

Introduction

The volume of applications being built on Machine Learning is large and growing. As of February 2017, over 2,300 investors had funded over 1,700 Machine Learning startups. Although a landscape of machine intelligence companies by industry exists, MI&S has not seen an application segmentation aligned by deployment environments. Such a model can be helpful in identifying which semiconductor architectures will be best suited for which applications and that are therefore likely to benefit from the growth in AI. (For background, MI&S published a summary of the various architectures used in Machine Learning and what to expect from the leading semiconductor vendors in 2017 here.)

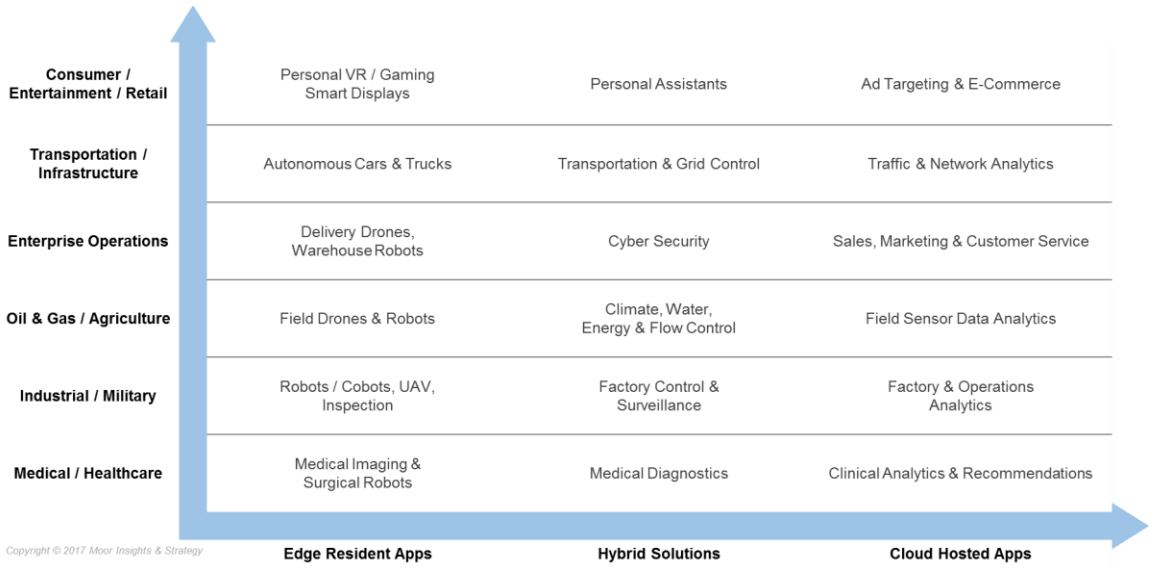
Consequently, this MI&S analysis portrays a new landscape for Machine Learning applications that are being developed and deployed in the **cloud**, at the **edge** of the network, and as interconnected (**hybrid**) applications or services. It then explores the various hardware architecture alternatives that may be appropriate to run these applications, while recognizing that the range of required hardware is as broad as the nature of the applications in each segment.

Machine Learning Applications

MI&S segments Machine Learning applications into 5 major use cases, aligning to the primary customer or beneficiary of the machine intelligence, each of which in turn are

sub-segmented by edge, cloud, and hybrid deployment environments. As a later section explores, the ideal hardware to support these applications spans a wide range of alternative architectures. The applications in Figure 1 all run some sort of pre-trained neural network, such as a deep neural network, a convolutional neural network, a recurrent neural network, *etc*. At this point, it is premature to size these market segments, but each represents a significant opportunity, and each is attracting a large number of solution developers, primarily in venture-capital backed startup firms.

FIGURE 1: A MACHINE LEARNING APPLICATION LANDSCAPE



(Source: Moor Insights & Strategy)

FROM THE EDGE TO THE CLOUD

Most AI services today reside in the cloud, where massive amounts of computational power can be harnessed to train the neural network and to handle the large volume of queries (called inference or scoring) in a cloud such as Microsoft, Facebook, Apple, or Google. Beyond the cloud, intelligent machines interact with people and objects through a device at the edge of the network. In consumer applications, the edge device may simply be a smartphone or tablet with access to intelligent services from the cloud, where the heavy lifting of neural network processing is performed. For example, consumer image classification services from Google or Facebook simply use mobile apps or browsers to provide end user input and output. In these cases, a desktop x86 or a mobile ARM CPU, perhaps augmented with additional neural network functions such as provided by the Qualcomm Snapdragon processor, will provide sufficient processing capability to meet the user's needs.

Beyond simple devices at the edge, vision- and voice-guided interfaces to intelligent systems require more local processing power and support for a growing range of sensors. This area is likely to experience significant growth, since natural language and image-based deep learning science has recently enabled precision approaching or even surpassing a trained human operator. In particular, the addition of vision processing greatly enhances the utility of the device, from autonomous vehicles to robots and drones to industrial controls and medical diagnostics. Here, latency can become a critical hardware requirement, as well as the need for additional processing beyond that of the neural network. These systems typically combine Machine Learning with computer vision, sensor fusion, and connectivity to other machines and to the cloud.

In these more advanced applications, the local device combines or "fuses" multiple input sources, performs local data processing, makes decisions about the data, then activates control systems in near real-time. These vision-guided autonomous systems demand significant local processing and low latencies, as the required reaction times preclude the luxury of passing the problem (data) to a cloud-based intelligence. Autonomous vehicles, mobile collaborative robots, collision-avoiding drones, and self-guided military unmanned aerial vehicles (UAVs) are some examples. An increasing number of these applications will not only require local intelligence but will be assisted by tightly linked cloud-based applications in a hybrid environment.

HYBRID APPLICATIONS, BETWEEN THE CLOUD & THE EDGE

Many applications will require an edge device to leverage cloud services to access more computational resources or data than can be maintained on the edge. An example of close cooperation between the cloud and the edge is the Google Home or Amazon Echo and Dot, as well as Alexa-based appliances. Using a combination of local and cloud based processing, the user is able conduct searches, order products, or launch a cloud-hosted "skill" from a library of over 10,000 applications. The ability to update those skills in the cloud keeps the technology fresh, as can be seen by the plethora of appliances at CES 2017, with Alexa or Home services providing the interfaces in the device and the smarts residing in the cloud.

A manufacturing example is a vision-guided bot or sensor in constant communication with an application in the cloud (public or, more likely, on-premises) to determine when a device or system is approaching the limits of its normal operating zone or reaching the conclusion of a process task. At that time, the cloud instructs the device to take appropriate action. Surveillance is another example, where local processing can identify

a potential threat, while cloud based processing can access databases and perform additional analytics to identify the specific threat and even individuals of concern.

Systems at the hybrid level can also act as an intermediary between the cloud, where the neural networks are trained, and the edge devices. The cloud can provide updated models to the device, and the device can provide experiential data to help the data scientists continually improve their models. In the case of FPGAs and reconfigurable SoCs which include FPGA technology at the edge, this reconfigurability also includes the ability to change the devices' hardware to take advantage of evolving algorithms, update sensor types and configurations, and update software algorithms. The resulting flexibility is a primary driver for FPGA and SoC adoption in these environments and is likely to create increased demand for these reconfigurable devices.

HYBRID HARDWARE FOR THE HYBRID ENVIRONMENT

This analysis focuses on the inference end of the machine learning workflow; it assumes that the training of the neural networks will be performed primarily on GPUs due to the computational (floating point) required. Each of the applications in this landscape has varying computational requirements when deployed in the cloud, edge devices, and tightly connected edge-cloud hybrid. The hardware requirements in each application are determined by the nature of the **data** that feeds into the system, the response time (**latency**) required for action, and the amount of synthesis or additional **processing** required. For example. processing live HD video at 32 frames per second requires far more compute power than processing text or simple still images. Low latencies are required for real-time applications, such as vision-guided drones or vehicles, while slower response times can be tolerated when dealing with people or many factory automation applications. Finally, beyond the job of performing inference using a trained neural network, synthesis may also be required to integrate multiple input streams (video, Lidar, Radar, *etc*.) and to implement a prescribed action. All of these factors must be carefully considered when selecting the right hardware for the job.

While the processing of data through the neural network often requires a dedicated accelerator (FPGA, GPU, DSP, or ASIC), the additional tasks mentioned above are best handled by a CPU (x86 or ARM), which can be programmed to execute more traditional algorithms developed in a traditional software lifecycle. To address this, FPGA vendors Xilinx and Intel (Altera) offer a range of SoCs with integrated CPU cores, programmable logic, and programmable sensor and connectivity interfaces. These "all programmable" devices bring together the accelerator required for DNNs and the CPU required for synthesis and implementing control policies with accelerated dataflow for sensor fusion

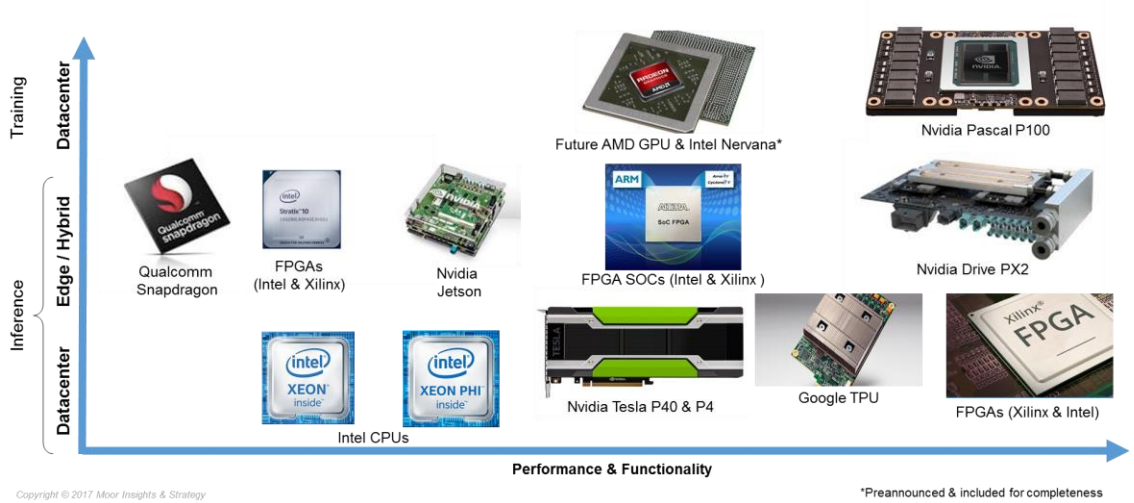
and real-time processing of multiple high-resolution video streams. In addition, Qualcomm offers inference on a flexible platform in the Snapdragon 835, consisting of CPU, GPU, and DSP for applications such as in-situ object recognition and security.

NVIDIA also addresses this area and offers a portfolio of platform-level approaches tailored for hybrid applications to augment their GPU and processing engines with sensors and connectivity. Specifically, the NVIDIA Tegra X1 targets this market at the chip level, while the NVIDIA Jetson and NVIDIA DrivePX2 platforms offer hybrid systems with ARM CPUs, sensor input I/O, and GPUs for the DNNs. These offerings target autonomous vehicles and applications such as drones, robotics, and even collaborative whiteboards. Hybrid systems can range from relatively light computational capacity to the heavy lifting required in real-time decision support systems involving high data rates, and MI&S expects they will increasingly appear across the landscape. Examples of these hybrid SOCs and systems include:

- NVIDIA DrivePX, Jetson TX1
- Intel Arria 10 SoC FPGA
- Xilinx Zynq All Programmable SoC
- Qualcomm Snapdragon 830

Figure 2 summarizes hardware architectures likely to be deployed across the landscape.

FIGURE 2: MACHINE LEARNING COMPUTE PLATFORMS



(Source: Moor Insights & Strategy)

HARDWARE ADVANTAGES & DISADVANTAGES FOR DEEP LEARNING

Previous sections provide context for hardware use cases across the landscape. This section looks at pros and cons of the alternatives. To be certain, one size does not fit all needs, and each alternative will find its appropriate use cases as the market develops.

GPUS IN THE CLOUD

GPUs excel at training in a cloud-based environment, where fine-grained floating-point accuracy is required to learn the weights in the neural network based on massive batches of (tagged) sample training data. Being well suited to training, advantages include ease-of-use, a well-established ecosystem, and an abundance of standardized libraries, frameworks, and support. GPUs can also be well suited for large-scale inference workloads and are available on many public clouds, including Google, IBM Softlayer, and Amazon AWS. (With the exception of IBM, however, these services do not yet offer the latest Pascal-based hardware, preferring to stick with the older Maxwell-based chips for the time being.) Disadvantages include high power consumption and lack of the ability to accommodate hardware changes as neural networks and algorithms evolve, particularly when compared to FPGAs.

CPUS IN THE CLOUD

The vast majority of inference queries in the cloud today are executing on servers with Intel Xeon CPUs. As the industry's ubiquitous server architecture, the Intel Xeon remains the chip of choice unless computational demands exceed the Xeon's capacity, which is increasingly the case for complex data types such as voice and video. Even here, the Intel Xeon Phi CPU is likely to do quite well, especially when it is upgraded later this year with 8-bit integer operations in the "Knights Mill" chip, providing a potentially significant improvement in query throughput.

FPGAS IN THE CLOUD

FPGAs excel at inference, where broad scale deployment requires the most compute efficiency in terms of performance-per-watt. They are also are reconfigurable, enabling leverage across a wide range of workloads and new evolving algorithms and neural networks. This reconfigurability is one of Microsoft's stated reasons for its decision to deploy FPGAs widely in its Azure and Bing properties. The state-of-the-art of inference is changing rapidly. Compression, pruning, and variable / limited precision (8-bit to 1-bit layers in the same network) techniques are examples where the algorithms are being refined and researched and may be broadly deployed. The major drawback to FPGAs

has been their difficulty to program. New software based development environments and development stacks are intended to address these shortcomings.

ASICs IN THE CLOUD

ASICs, such as Google TPUs, can provide the highest compute efficiency of all alternatives for targeted applications. They typically lack the ease-of-use of GPUs and the reconfigurability of FPGAs. They are also costly to create and might become obsolete quickly as algorithmic advances become available. While Google is the only company known to have developed an ASIC for Machine Learning, several startups are developing their own ASIC for this market over the next few years.

TYPICAL EMBEDDED SOCS & EMBEDDED GPUS

For edge applications, typical SoCs and embedded GPUs (with processor cores) such as NXP i.MX, Qualcomm Snapdragon, and Nvidia Tegra X1 can be relatively easy to program and offer low cost and reasonable throughput. However, because they must continuously access external memory, they are typically limited in the latency they can support. They also typically have limited sensor and connectivity interfaces and have fixed hardware, which limits their ability to be upgraded as algorithms, networks, sensors, and interface requirements change.

EMBEDDED RECONFIGURABLE SOCS & FPGAS

Embedded “all reconfigurable” SOCs and FPGAs from Intel and Xilinx provide the ability to create customized dataflow engines and accelerators without the need to continuously access external memory, enabling high throughput and low latency sensor fusion, vision processing, and Machine Learning, along with full deterministic control. Traditionally, these devices have been relatively difficult to program. New software-based development tools and development stacks are being introduced to simplify programming and enable new users who have limited hardware expertise.

CONCLUSIONS

Applications demand varying computational capacity depending on where they reside: in the cloud, edge devices, or in a hybrid environment. The computational requirements vary with the type of data being analyzed and with the factors dictated by the AI environment where it operates. Many edge device environments require more complex computational platforms with a mix of tightly coupled CPUs and accelerators (FPGAs, DSPs, or GPUs in new types of SoCs and platforms). Such platforms are required for

the AI to determine the proper course of action in near real-time while maintaining connectivity to the cloud for updated models and to provide experiential data to continually improve the cloud-based training of the underlying neural networks. These hybrid devices are likely to see significant adoptions in intelligent edge applications.

The hardware underlying the advances in Machine Learning is evolving rapidly along with the Machine Learning frameworks and libraries. In such a dynamic environment, companies deploying these systems need to plan for the future carefully to adapt to new approaches without resorting to a rip-and-replace upgrade.

IMPORTANT INFORMATION ABOUT THIS PAPER

AUTHOR
Karl Freund, Senior Analyst at Moor Insights & Strategy

PUBLISHER
Patrick Moorhead, Founder, President, & Principal Analyst at Moor Insights & Strategy

EDITOR / DESIGN
Scott McCutcheon, Director of Research at Moor Insights & Strategy

INQUIRIES
Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

CITATIONS
This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

LICENSING

DISCLOSURES
This paper was commissioned by Xilinx, Inc. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER
The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.