Machine Learning Moves to the Edge

By Linley Gwennap
Principal Analyst

April 2020

www.linleygroup.com

Machine Learning Moves to the Edge

By Linley Gwennap, Principal Analyst, The Linley Group

Embedded systems at the edge of the network are at the beginning of a revolution. These devices will increasingly use machine learning (ML) to deliver new capabilities. Even as ML becomes more important, these new capabilities must be added without breaking the BOM cost and power budgets of existing products. The best option is an SoC that includes the CPU, standard interfaces, and an efficient ML accelerator on a single chip, minimizing cost, power, and board area. SiMa.ai is a new company developing a product to fit these requirements. SiMa.ai sponsored this white paper, but the opinions and analysis are those of the author. Trademark names are used in an editorial fashion and are the property of their respective owners.

Introduction

Machine learning, often referred to as AI, is the most powerful trend in the technology industry. By employing deep neural networks instead of hand-coded software, users can quickly enable a processor-based system to recognize images or words and even respond to them. But neural networks can be very compute intensive and require custom hardware to improve power efficiency. For these reasons, most machine learning (ML) today runs on cloud-based servers. For example, web-based surveillance cameras continuously transmit video to the cloud, which employs a neural network to recognize unusual activity.

This process simplifies the camera's software, but it has several problems. Sending video to the cloud raises privacy concerns, particularly for R&D labs and other secure facilities. Transmitting a continuous high-definition video stream ties up the user's network, the cloud-service provider's network, and the infrastructure in between. Cloud-based services are unusable any time the network connection stalls or fails. And for cars, drones, and robots, the latency to the cloud is far too slow for remote decision-making. Thus, moving ML from the cloud to the edge is valuable for many camera-based applications and essential for autonomous devices.

As these examples shows, the opportunities for ML at the edge of the network are broad. Governments from Beijing to London and beyond are deploying surveillance cameras at a rapid rate: 200 million over three years in China alone. Businesses are also using smart cameras not just for security but to count customers and monitor their behavior. Smart cities use cameras to monitor traffic, parking, and other infrastructure. Factories and warehouses increasingly rely on robotic systems. Cars and drones must avoid crashing even if their drivers make mistakes. Medical equipment can employ ML to interpret scans or monitor patients. As Figure 1 shows, these markets represent a large and diverse opportunity for semiconductor vendors.

These applications are already starting to adopt edge-based ML. Many new surveillance cameras in China and elsewhere use a neural network to locate and extract multiple faces from an image (e.g., a busy sidewalk) and send them to the cloud for facial recognition. The SimCam, a consumer product, can identify family members using in-device

facial recognition and send an alert if a stranger enters the home. Most new cars today have features such as automatic emergency braking (AEB) that rely on cameras to detect road hazards. High-end consumer drones, such as Skydio, can autonomously track a moving person to capture action video while avoiding obstacles. As the cost and power of implementing ML declines, these features will increasingly appear in new products.

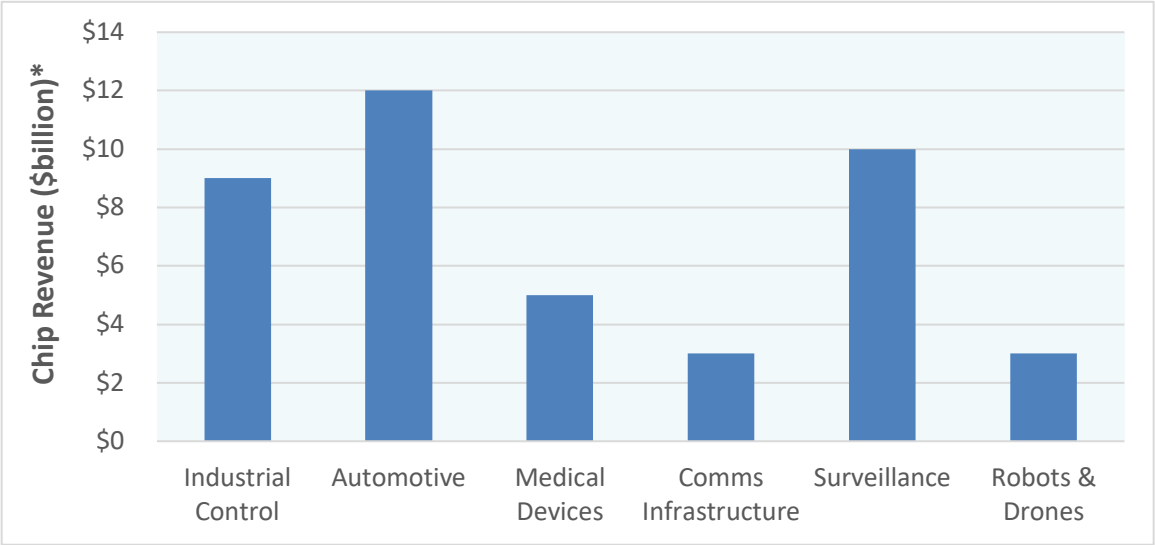

Figure 1. Edge-device markets for machine learning. ML addresses a broad range of end applications representing billions of dollars of semiconductor spending. (Source: Tractica, others)

Technology Challenges

A common theme of these emerging applications is computer vision. Because of the tremendous volume of smartphones, high-quality camera sensors have become very inexpensive. Even low-cost consumer devices can now add one or more cameras, allowing them to better interact with their surroundings and their users. Computer vision enables user identification (through facial recognition), object identification, and location identification, among other capabilities. For surveillance applications, the trick is to identify the tiny fraction of a video stream that could be anomalous behavior.

Since 2012, deep neural networks have been the winners of the ImageNet competition, defeating hand-coded programs. Neural networks can be trained to recognize specific faces or types of objects simply by feeding them labeled images; no additional programming is required. This machine-learning approach simplifies the deployment of computer-vision functions and enables more sophisticated capabilities, such as the aforementioned emergency braking for cars and user tracking for drones. Researchers continue to develop more complex neural networks that provide greater accuracy and handle larger images, but these larger networks require more compute performance.

Although ML can perform much more sophisticated analyses than standard software, it also requires far more calculations. Performing these calculations on a standard CPU, or even a more-specialized GPU, is very taxing. Even simple neural networks can push typical edge processors to their maximum power; more complex networks require high-end processors that burn hundreds of watts. Researchers continue to develop new

networks and algorithms that are too power-hungry to run on today's processors. Thus, the market is primed for a new, more efficient ML architecture.

While machine learning is becoming an important capability, it must be added to an existing system rather than designed in isolation. For example, an autonomous drone must still manage its flight system and camera(s) while recording or transmitting video. Existing products often have a large base of traditional code that must be augmented, not replaced, with new ML functions. This code may represent dozens or even hundreds of engineer-years of effort, so it is not easily replaced. Compatibility with legacy code is thus a paramount requirement for many embedded designs.

For most of these systems, the legacy code is developed for ARM-compatible CPUs. To maintain compatibility, the simplest approach is thus to run ML functions on the existing ARM CPU. This CPU, however, is often underpowered for the intense computing needed for many neural networks. Even if the CPU can handle the task, it is far less power efficient than an optimized ML accelerator. Thus, adding ML to an existing design often reduces battery life or requires a larger, heavier battery. Alternatively, designers can add a separate ML chip to the system, but this approach adds cost and board area. The most efficient method is to employ an SoC that combines an ARM CPU for legacy code with an integrated ML accelerator that can handle sophisticated neural networks.

Application Requirements

Most surveillance cameras are line powered, but their power draw is still constrained. Many industrial and government cameras employ Ethernet networking, which can also provide up to 13W of power. Even cameras that connect directly to electrical wiring are often kept below 10–15W to reduce the cost of continuous operation. To add machine learning, these surveillance cameras can't use a typical GPU that requires 75W or more; they need an ML chip that consumes just a few watts. Within this power budget, simple segmentation (e.g., locating faces) requires about one trillion operations per second (TOPS) for a 720p image. Running a more sophisticated image-recognition model such as YOLOv2 at 1080p and 30fps requires about 12 TOPS.

Traditional industrial robots perform repetitive work on an assembly line, but factories now seek to automate robots that can locate and identify components for assembly, or identify and pack items for shipping. These and other sophisticated tasks require computer vision, often using multiple cameras for depth perception. They also require controlling a robotic arm to manipulate a variety of objects that could be in different locations and orientations. Neural networks for these capabilities can range from 10–50 TOPS. Some of these robots are mobile, so they can fetch items or operate at different work stations; for these battery-operated devices, power consumption for machine learning should be below 10W. Even for stationary line-powered robots, ML processing at less than 20W reduces the size and operating cost of the system.

Advanced driver assistance systems (ADAS) typically employ a front-facing camera on the rear-view mirror. This location requires small, low-power systems that consume less than 10W. ADAS functions such as emergency braking (AEB) require relatively low

performance of 1–5 TOPS. The automotive industry is now developing systems for hands-free autonomous driving that require far more performance. Because these systems have not yet been deployed, the requirement is unknown but likely to be at least 100–500 TOPS, depending on the types of driving conditions and locations that are supported. Prototype autonomous systems burn hundreds or even thousands of watts, but this power draw isn't deployable in consumer vehicles because of the heat generated and the reduction in mileage/range for gas/electric vehicles. Although some commercial vehicles (such as robotaxis) will deploy at higher power, autonomous driving systems must reach 100W or below to appear in mainstream consumer vehicles.

Medical applications for machine learning include interpreting medical scans and monitoring patent activity and even vital signs using cameras. Because these tasks require more sophisticated analysis than simple video surveillance, they can require from 10 TOPS to 50 TOPS or more. Power budgets in health care are more relaxed than in consumer equipment but should still be less than 50W.

SiMa.ai Solution

SiMa.ai was founded in early 2019 to solve these problems. *Sima* means *edge* in Sanskrit, and the company targets edge applications that require high-performance vision processing without the high power and cost of GPU-based accelerators. To do so, the company has developed the MLSoC architecture, which encapsulates a single-chip solution for embedded systems that require machine-learning capability.

As Figure 2 shows, the MLSoC includes a standard ARM-compatible CPU as well as standard interfaces such as DDR DRAM, Ethernet, video (camera) connections, and a debug port. The MLSoC includes a security block that performs encryption and other functions to deter hackers and other threats. It also features a complete safety block that enables designs that meet ISO 26262 and other important automotive standards. A network-on-chip efficiently connects all these subsystems.

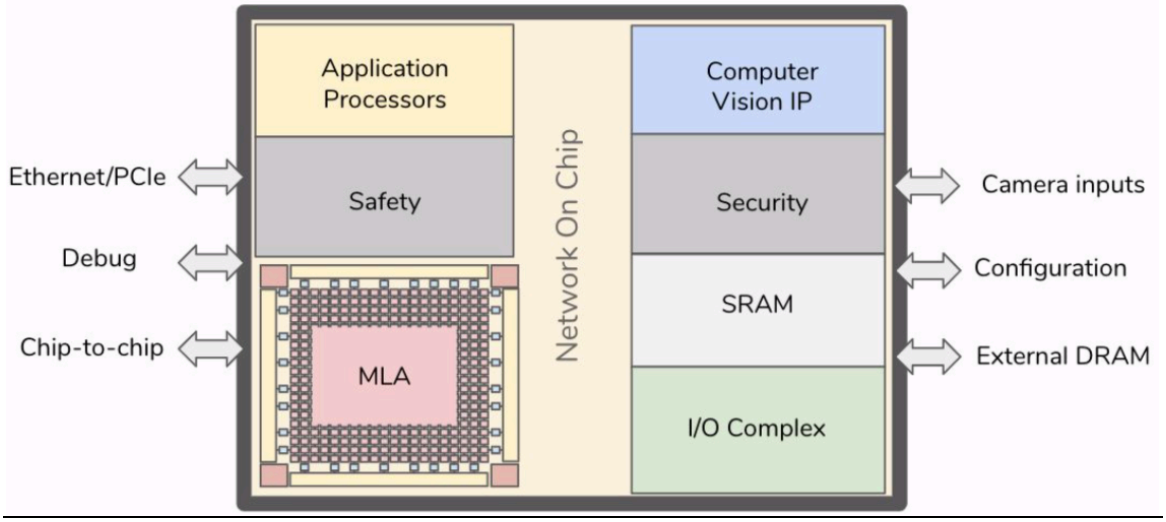

Figure 2. SiMa.ai MLSoC block diagram. The MLSoC combines a standard Arm-compatible SoC with SiMa.ai's custom ML accelerator in a single chip.

The chip's unique feature is the SiMa.ai ML accelerator (MLA). This custom block has enough performance to run complex neural networks, but it consumes much less power than traditional GPUs. The MLA employs a new architecture that shifts complexity from the hardware to the software, thus reducing power consumption. The design is optimized for the matrix computations that are common in many types of neural networks, improving efficiency over CPUs (which are optimized for scalar computation) and GPUs (vector computation). The MLA architecture can scale from 50 TOPS to 200 TOPS and beyond, enough for even the most demanding edge applications. The company is also developing a tool chain to simplify porting ML applications to the chip.

This heterogeneous design enables the MLSoC to easily fit into many camera-based embedded systems. Its ARM CPU can run legacy code to control the system. The security and safety blocks can offload other common system functions, although some software change may be required. The standard interfaces allow designers to connect the chip to the most common type of memory, the most common networks, and the most common camera sensors. Although replacing a standard SoC with the MLSoC requires changing the board design, the new board typically maintains a similar size and power, avoiding changes to the rest of the system.

The ML accelerator enables customers to add machine learning without breaking their power budget. In fact, the MLSoC is designed to deliver industry-leading power efficiency of 10 TOPS per watt. By comparison, leading GPUs offer about 1 TOPS per watt, and special-purpose accelerators can achieve as much as 5.2 TOPS/W.

This incredible efficiency translates directly to strong performance on real neural networks. For example, SiMa.ai expects the MLSoC to achieve 2,280 images per second (IPS) on ResNet-50 inference when running at a batch size of one (batch=1), which is typical for real-time video analysis. Nvidia rates its Xavier processor at 1,390 IPS on the same test. But while Xavier consumes 29W at this speed, the MLSoC will achieve its strong performance while using about 4W. In other words, the MLSoC is rated at 570 IPS/W, versus just 48 IPS/W for Nvidia's best edge processor. The leading accelerators top out at 395 IPS/W.

Conclusion

Embedded systems at the edge of the network are at the beginning of a machine-learning revolution. These devices will increasingly add ML functions to deliver new capabilities. Some of these devices, such as smart speakers and security cameras, already deliver new capabilities while relying on cloud-based ML processing. But moving this processing to the edge device improves privacy and reliability while reducing cost and network bandwidth. Because of long latency to the cloud, in-device processing is essential for autonomous operation of automobiles, drones, and robots.

Even as ML becomes more important, these new capabilities must be added without breaking the BOM cost and power budgets of existing products. Many of these products have strict power limits, and consumer products in particular must stay within certain price points. A standard CPU can't meet the performance requirements of many ML applications, so an ML accelerator is essential for many new systems. This accelerator

must be small and power efficient to fit into existing designs. Yet the accelerator can't run legacy code, so it must be paired with a standard CPU and standard interfaces. The best option is an SoC that includes the CPU, standard interfaces, and an efficient ML accelerator on a single chip, minimizing cost, power, and board area.

SiMa.ai is a new company developing a product to fit these requirements. Its MLSoC provides a custom ML accelerator that is rated at 10x better power efficiency than leading GPU-based designs and 2-3x better efficiency than the best accelerators available today. The MLSoC also includes a popular ARM-compatible CPU for legacy code and standard DRAM, Ethernet, and camera interfaces to simplify system design. Companies developing camera-based edge systems should consider the MLSoC for adding intelligence to their next design.

Linley Gwennap is principal analyst at The Linley Group and editor-in-chief of Microprocessor Report. *The Linley Group offers the most-comprehensive analysis of microprocessors and SoC design. We analyze not only the business strategy but also the internal technology. Our in-depth articles cover topics including embedded processors, mobile processors, server processors, AI accelerators, IoT processors, processor-IP cores, and Ethernet chips. For more information, see our website at www.linleygroup.com.*