Enabling AI Vision at the Edge

Introduction

Computer vision has made tremendous advances in the last several years due to the proliferation of AI technology. The intersection of big data and massive parallel computing changed the way in which machines are programmed to understand unstructured 2D and 3D data, such as video feeds from cameras. Instead of writing a set of rules, programmers use deep convolutional neural network algorithms to enable a machine to generalize a solution from a large, labelled dataset. This new technique, AI vision, has enabled machines to reach super-human capabilities in accurately identifying objects in an image. AI vision makes the camera one of many intelligent sensors.

AI vision is now deployed in many new applications such as

- Autonomous vehicles
- Smart cities and agriculture
- Industrial and warehouse robotics
- Delivery drones and robotic cleaners
- Augmented reality
- Smart retail and smart home

Traditionally, much of the AI vision processing was performed in the cloud with massive parallel compute capabilities. However, with the mass deployment of the AI vision, streaming video from a camera to the cloud for processing exceeds the available network bandwidth. A raw 1920x1080 camera operating at 30 FPS will generate ~190 megabytes per second of data. With billions of devices deployed, even with H.265 compression, over 5 petabytes per second of data will be required. In addition, privacy concerns and round-trip network latencies are moving the AI processing to the edge where the data is generated. Custom SoCs are required to meet the power constraints of AI vision at the edge.

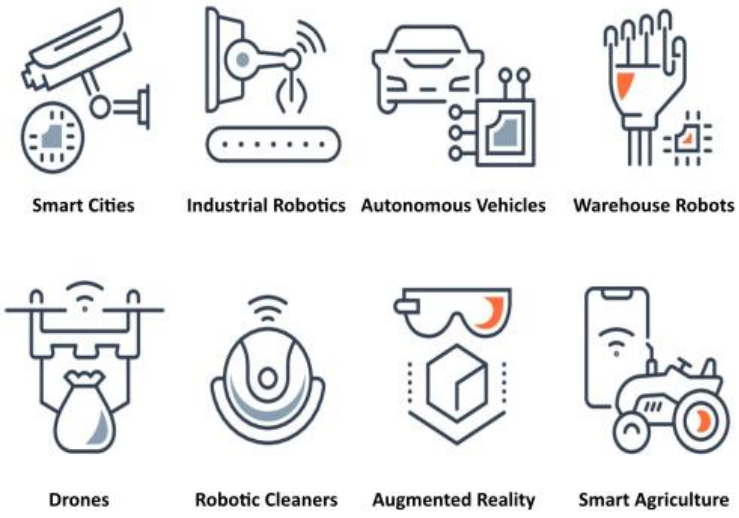

Figure 1. AI vision applications

AI Drives New Growth in ASICs

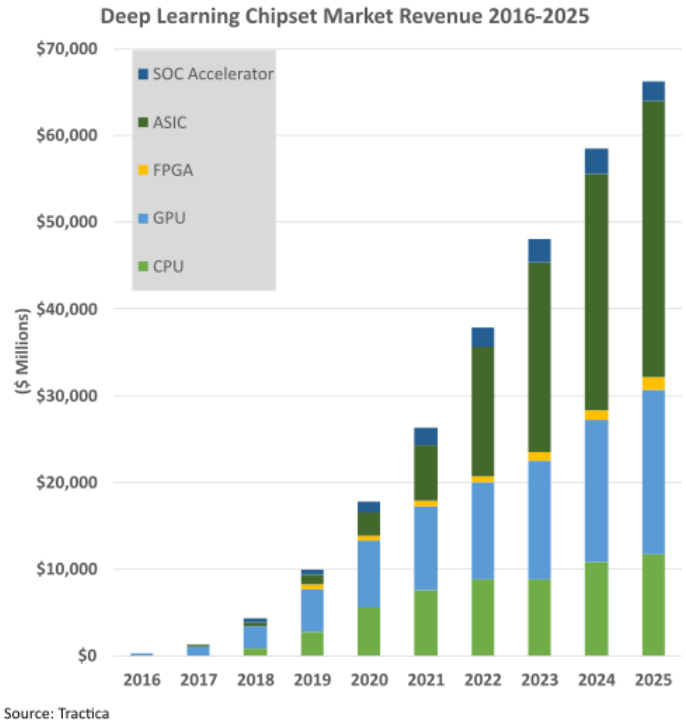

Figure 2. AI market to exceed $66.3 billion in 2025

Due to its superior performance relative to traditional machine learning techniques, AI is being deployed in all markets from consumer to industrial to automotive applications. This is spurring growth in new AI-enabled hardware in both the cloud and the edge. More specifically, as shown in Figure 2, the total AI market will grow to $66.3 billion in 2025, representing at 60% CAGR [1].

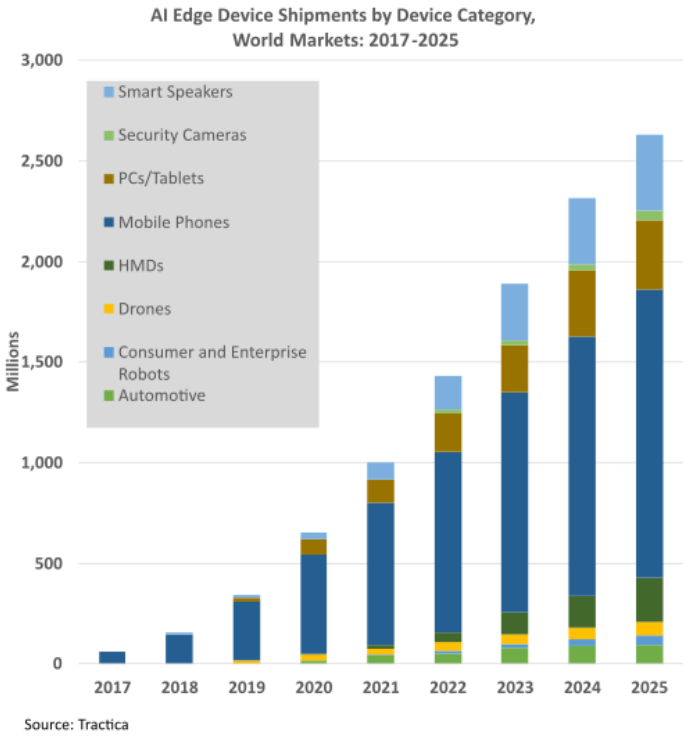

Figure 3. AI edge expands from mobile into embedded vision

Today, many of the hardware run AI on general purpose CPUs and GPUs. However, with increasing performance requirements, more specialized SoCs are being developed with custom AI accelerators to meet these requirements within strict power constraints. By 2025, 60% of the AI edge training and inferencing hardware will use custom ASICs and accelerators [1].

AI vision at the edge was first deployed in mobile smartphones. As shown in Figure 3, the adoption of the technology is expanding to other embedded applications, such as robotics, security cameras, and automotive. By 2025, the number edge devices with AI Vision capabilities will exceed 2.6 billion units [2].

From Camera to Metadata over the Network

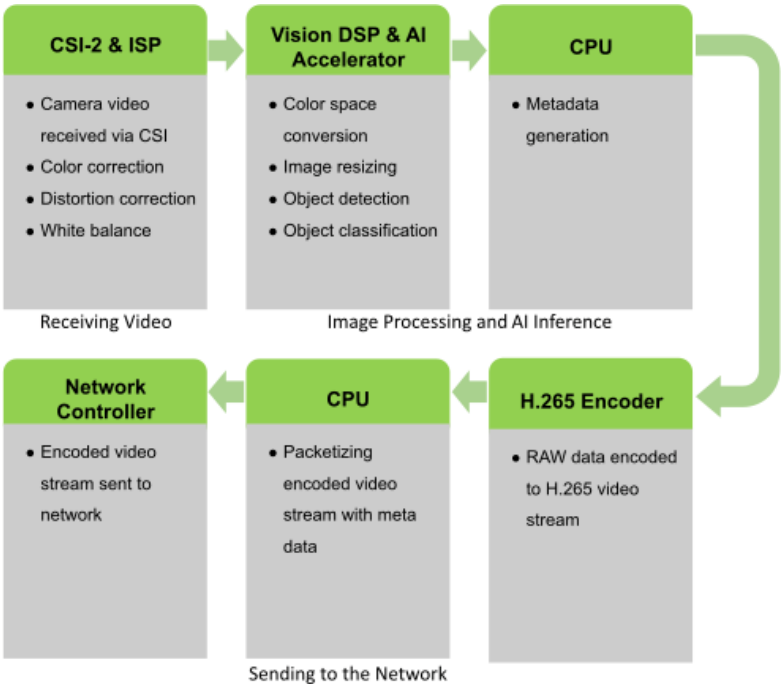

Figure 4. Typical data flow from camera to the network

A typical AI vision SoC must analyze a video stream from one or more cameras and make real-time inferences about a scene, such as the types of objects and intent of objects. From these inferences, along with data from other sensors, the AI vision SoC needs to make decisions, such as controlling the direction of a robot, and send the metadata to the cloud.

Figure 4 shows an example of a data flow within an AI vision SoC. A camera sends the raw video stream to the AI vision SoC through the CSI interface. The incoming raw video stream is pre-processed in the ISP (color and distortion correction, white balance, etc.). The vision DSP takes each frame of the video and performs color space conversion and image resizing. The vision DSP and the AI accelerator run multiple neural networks on the resized frame to extract the pertinent information from the frame. The CPU generates the metadata and may embed the metadata in the video stream. The video stream is then compressed with the H.265 encoder. The CPU then packetizes the compressed stream and metadata, and the network controller sends the packets over the network back to the cloud.

Speeding Development with OpenFive's AI Vision Platform

Building and deploying a custom AI vision SoC with your custom accelerators used to require significant time and expertise. Traditionally, you would need to source all the IPs, integrate them with your custom accelerators, tapeout, and productize it with a software development kit. Meanwhile, you are racing with your competitors to be first to market.

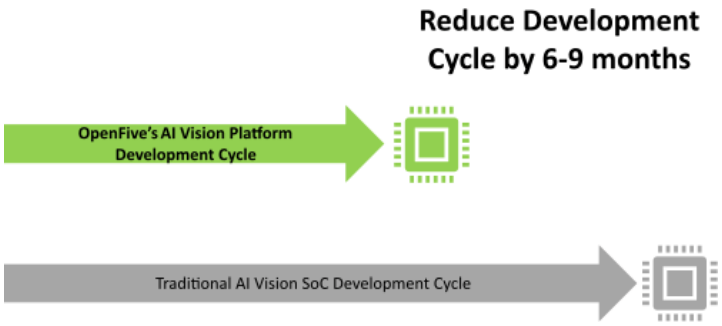

Figure 5. Reduce development cycle by 6-9 months

OpenFive's AI Vision platform is the solution to speed up deployment of your custom AI vision SoCs. OpenFive's AI Vision platform is composed of multiple customizable subsystems so that you can focus your key differentiator for the end application.

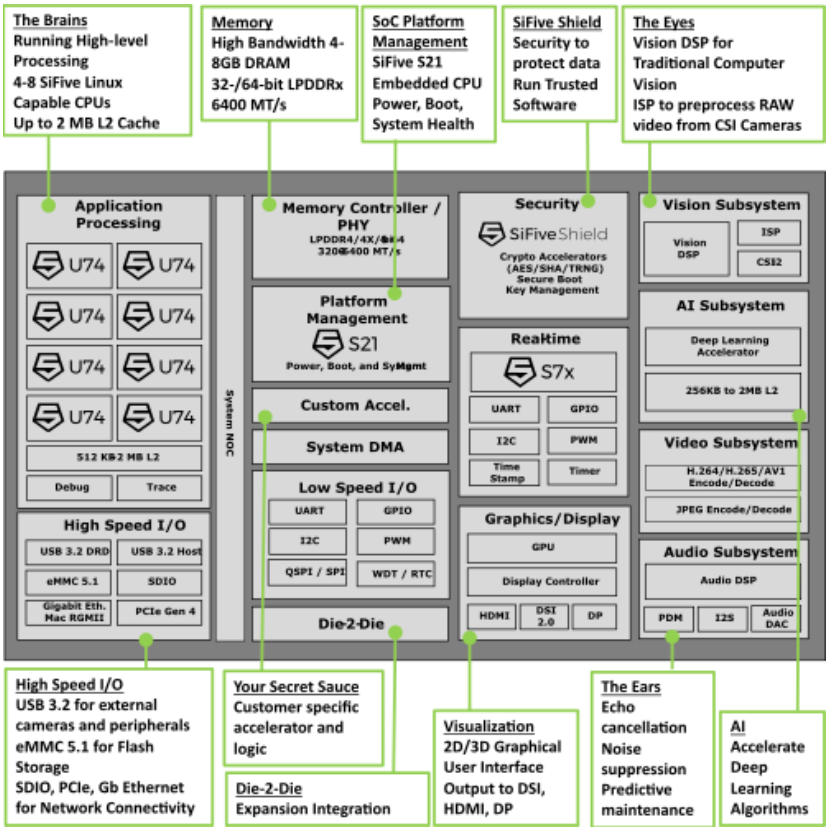

Figure 6. OpenFive's AI Vision platform

The application processing subsystem of the AI Vision platform enables running high-level applications, such as navigation, networking, and graphical user interfaces. It is powered by the SiFive's multicore Linux-capable CPUs and is configurable from four to eight cores. The vast Linux software ecosystem abstracts all the low-level hardware and provides APIs required to build targeted applications quickly.

The vision subsystem with the ISP processes the incoming video streams from cameras connected to CSI or USB. The vision DSP runs computer vision algorithms, like SLAM, on the video streams. In addition, video can be decoded and encoded with the hardware accelerated H.264, H.265, AV1, and JPEG codecs.

The real-time processors, like SiFive's S7, with tightly integrated peripherals provide real-time sensing and control. High-level commands are typically sent from the application processing subsystem for the real-time processors to manage motor control and actuation.

The AI Vision platform supports LPDDR4/4x/5 for gigabytes of high-speed memory and can boot from Quad SPI NOR FLASH or eMMC 5.1. The platform has a variety of I/O—I2C, UART, SPI, SDIO, and GPIOs—to interface with off-chip peripherals, such as sensors and wireless networking. For high-speed connectivity, USB 3.2 and PCI Express can connect to a variety of modems, external cameras, and peripherals.

The entire AI Vision platform is secured by SiFive Shield, which provides crypto-hardware to accelerate AES, SHA, TRNG, and more. SiFive Shield's Root-of-Trust and secure boot mechanism ensure that only trusted software is run on the AI Vision platform.

If a graphical user interface is required, the AI Vision platform has a visualization subsystem with a 2D/3D GPU for rendering graphic-rich human interfaces and can drive display panels through DSI, HDMI, or DisplayPort interfaces.

Because audio is also important to sensing at the edge, the audio subsystem has integrated I2S interfaces, digital microphone inputs, and audio DACs. Additionally, the audio DSP can be used for echo cancellation and noise suppression on the incoming audio streams.

The AI subsystem has hardware acceleration to run object detection, object recognition, and segmentation on the video streams using the deep convolutional neural networks, such as YoloV3, Resnet50, and MobileNet. This AI subsystem can also be augmented or replaced with your own custom AI accelerators.

Besides custom AI accelerators, OpenFive can integrate your other custom accelerators. Performance can be further enhanced by adding additional chiplets using the die-2-die interface.

SDK and FPGA Emulation

A major benefit of OpenFive's AI Vision platform is leveraging a base software development kit and an FPGA emulation platform to start software development early. With drivers and subsystems pre-integrated, you can focus on building and testing your application while the SoC is being designed. Enhancement and optimizations identified early can be incorporated into the SoC before tapeout.

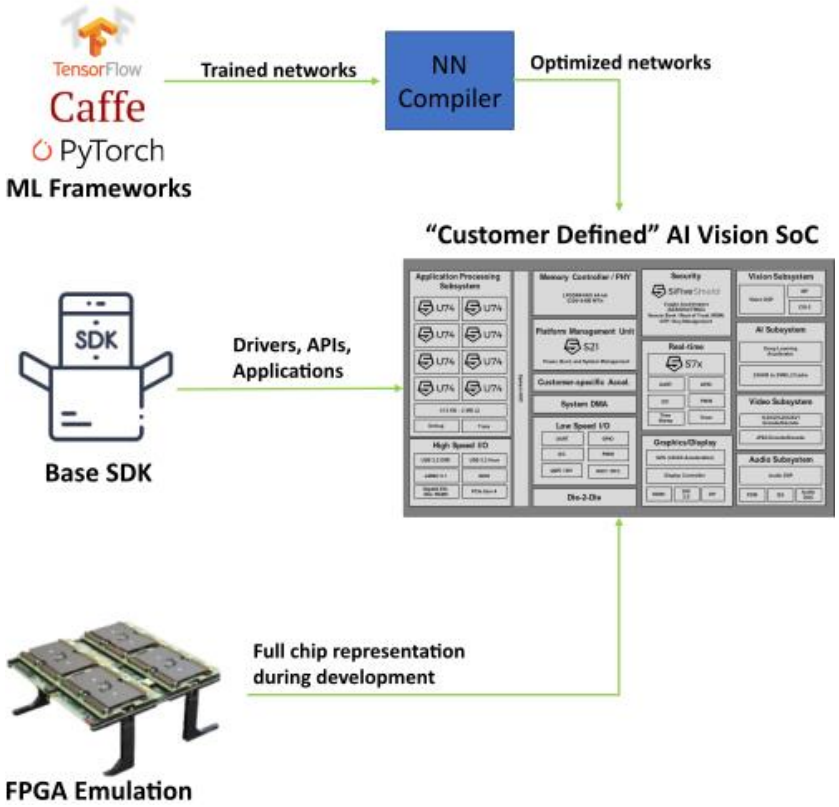

Figure 7. Jump start software and algorithm development

Additionally, if you have trained your AI model in one of the popular machine learning frameworks, like TensorFlow, Caffe, and PyTorch, you can leverage the neural network compiler from our ecosystem partner to load onto their AI accelerator in the OpenFive AI Vision platform.

Start Customizing Your AI Vision SoC

OpenFive's AI Vision platform speeds up deployment of AI vision SoCs for the rapidly evolving and growing edge AI market. OpenFive's AI Vision platform's pre-integrated hardware and software mean you can focus on your key differentiators and get to market as quickly as possible.

References

1. P. Clarke. "AI chip market set to rival that of microcontrollers," eeNews Europe. 19 September 2018 [Online]. https://www.eenewseurope.com/news/ai-chip-market-set-rival-microcontrollers-0.
2. "Artificial Intelligence Edge Device Shipments to Reach 2.6 Billion Units Annually by 2025", Edge Ai + Vision Alliance. 24 September 2018 [Online]. https://www.edge-ai-vision.com/2018/09/artificial-intelligence-edge-device-shipments-to-reach-2-6-billion-units-annually-by-2025.

Authors:

David Lee

Director of Product Management, SoC and Systems, and AI Vision Platforms, OpenFive

David Lee is the Director of Product Management, SoC and Systems, and AI Vision Platforms at OpenFive. Previously, he was at Marvell and NVIDIA, where he held various positions in engineering, architecture and technical marketing for multiple generations of Marvell's Armada and NVIDIA's Tegra SoCs. He has been designing and architecting systems for 15+ years, including NVIDIA's Jetson TX1, SiFive's HiFive Unleashed and Unmatched boards and, more recently, BBC's Dr. Who Inventor Coding Kit. Additionally, he has architected mobile phones, tablets, and self-driving car computers. David has an MS in Electrical Engineering from UCLA, and a BS in Electrical Engineering and Computer Science from UC Berkeley.

Christopher Moezzi

VP & GM of AI Solutions, OpenFive

As VP & GM of AI Solutions, Chris manages OpenFive's AI Platform development for embedded vision and edge AI inference applications. Chris brings over 20+ years of semiconductor industry experience, with a proven track record of managing large-scale SoC and ASIC businesses at top-tier companies such as Broadcom, Marvell, Cavium, and Faraday. As a Vice President at Marvell, Chris was responsible for the company's LiquidIO® Data Center/Cloud SmartNIC and Ethernet Server adapter business. At Broadcom, he grew the StrataGX™ Embedded ARM® SoC business to record revenues with a portfolio of MCUs and Application Processors. Chris holds a Master of Science degree in Electrical Engineering (MSEE) from Southern Methodist University, and an MBA from Northeastern University.