AI and ML Offer a Completely New Way to Solve Problems

Interview with Markus Levy
Director of AI and Machine Learning Technologies, NXP Semiconductors

Question: Are AI and machine learning buzzwords or is the hype behind AI/ML real?

Markus: For sure the AI and/or ML buzzwords are hyped, but it's more real than the hype that surrounded IoT 5 years ago. This is not meant to undermine the value of IoT, because it has certainly led to the development of many cool new application areas, but in many ways IoT was an evolution of embedded. On the other hand, AI/ML reflects a 'market segment' that provides an entire new way to solve problems. Remember that machine learning in the form of neural networks and classical algorithms has been around for decades. However, regarding deep neural networks, no one knew how to practically train these and it was impractical using commodity hardware, so they remained relatively dormant until recent years. Now the amount of research has exploded, and so has the availability of open source and proprietary options to train, optimize, and deploy machine learning models. Today we see many more companies investing resources and money into AI/ML development, and this is just the tip of iceberg. However, as with any cutting-edge technology where people 'smell' money, there will be winners and losers.

Question: How do you see AI proliferating into mainstream technology?

Markus: First you should note that for AI/ML to proliferate across all or many application areas and become mainstream technology, it must be simplified to a level that models can be trained and inference engines can be developed & deployed with a python script or pull-down menu, rather than burdening the developer with having to create complex mathematical algorithms. Programming for AI/ML is often referred to as Software 2.0. It's not about using traditional programming methods, instead it's about utilizing existing neural networks and frameworks and classical ML libraries. To support this, the effort is more about determining weights and parameters (e.g. typically known as training models).

"In the medium and long term, we want to integrate some kind of AI acceleration into many of our components. It even goes so far as to think about special hardware blocks on the low-end MCUs."

– Markus Levy

Related Article: AI-ML and Thor96 single-board computers: specially built for AI, IoT, visual display and more

Regarding the particular benefits of AI, fundamentally this is an area where the data is gold. In other words, AI/ML is more or less useless unless the application developer can collect, and subsequently label, data that is used for training the models. Assuming that the application developer is able to train the model, the benefits of AI can be applied in a wide range of applications targeted at vision, voice, and anomaly detection.

Question: Where does NXP stand in AI/ML leadership?

Markus: NXP is recognized as a leader in the AI environment because we address the AI applications directly from a software and complementary hardware perspective. In other words, the first ingredient to successful deployment of AI is having a scalable portfolio of hardware platforms to give developers a choice for performance, power, and price. The ability to run an ML algorithm on a low-end MCU as well as a high-end applications processor is driven by the acceptable inference latency and the memory footprint. Software is the second ingredient required for AI deployment, and for this, NXP has chosen a path that provides a variety of open source options, we refer to this as the eIQ machine learning software development environment. It's become obvious that proprietary solutions for machine learning deployment will never be able to keep up with the rapid progress we see happening with open source. Therefore, our mission is to enable open source options, such as TensorFlow Lite, GLOW, Arm NN, and others, and apply device-specific optimizations as needed to achieve more competitive results.

Question: What are the applications covered in NXP's AI/ML enablement environment?

Markus: As I mentioned earlier, we focus on vision, voice, and anomaly detection – of course this is very broad and at a high level represents most of applications residing at the edge. Vision can be subdivided into applications such as face recognition and object detection and recognition and can even cross over into the anomaly detection domain. Voice encompasses keyword detection for Alexa-like applications, as well as limited natural language processing (limited because of the restrictions on memory capacity on edge devices). And anomaly detection spans applications such as predictive maintenance and monitoring a wide variety of sensors to detection normal and abnormal circumstances.

Question: What are the features of NXP chips that make it uniquely suited for AI/ML applications?

Markus: A scalable portfolio will lend itself well to these AI applications. It all comes down to how much performance is required (in other words, the inference latency time) and the amount of memory available. People will say that they want more performance and fancy ML accelerators but cost is always the overriding factor. For that matter, an NXP product like the i.MX RT600 (with integrated DSP) will be an excellent choice for voice and vision applications. The integrated DSP is used for keyword detection, and it allows most of the device to remain in low-power mode until the appropriate keyword is detected. At the other end of the spectrum, NXP offers i.MX 8 and Layerscape family devices, with integrated GPUs, multicore CPU cores, and DSP, giving the opportunity to perform heterogeneous computing and multiple machine learning algorithms in parallel. Next generation devices will include dedicated machine learning accelerators, yielding unprecedented performance levels.

Question: What is NXP's eIQ machine learning enablement?

Markus: At the base level, eIQ is a collection of open source technologies to deploy machine learning applications. These technologies include runtime engines such as TensorFlow and TensorFlow Lite, it includes network parsers and dedicated inference engines such as Arm NN, and support for libraries such as OpenCV. While all these technologies can be obtained from sources such as Github, NXP's eIQ makes these easier to use, provides detailed documentation, and integrates these into our development environments (MCUXpresso for the MCUs, and Yocto/Linux for the i.MX application processors).

Question: What is planned next for NXP hardware to support AI/ML applications?

Markus: The medium- and long-term goals are to include some form of AI acceleration in many of our devices, even specialized hardware blocks on the lowest-end MCUs.

Coming Soon - AI and ML Single-Board Computers based on NXP i.MX 8 Application Processors

AI-ML Board based on NXP i.MX 8X SoC and the Thor96 Board based on NXP i.MX 8M SoC. Contact Arrow to reserve your board.

In Person

North America
+1 855 326 4757

Europe, Middle East and Africa
+44 20 3936 5486

Asia-Pacific
+86 400 920 0628

Online

arrow.com/iot

Five Years Out