*Technical Article*
Bringing the Next Evolution of Machine Learning to the Edge

Mark Nadeski
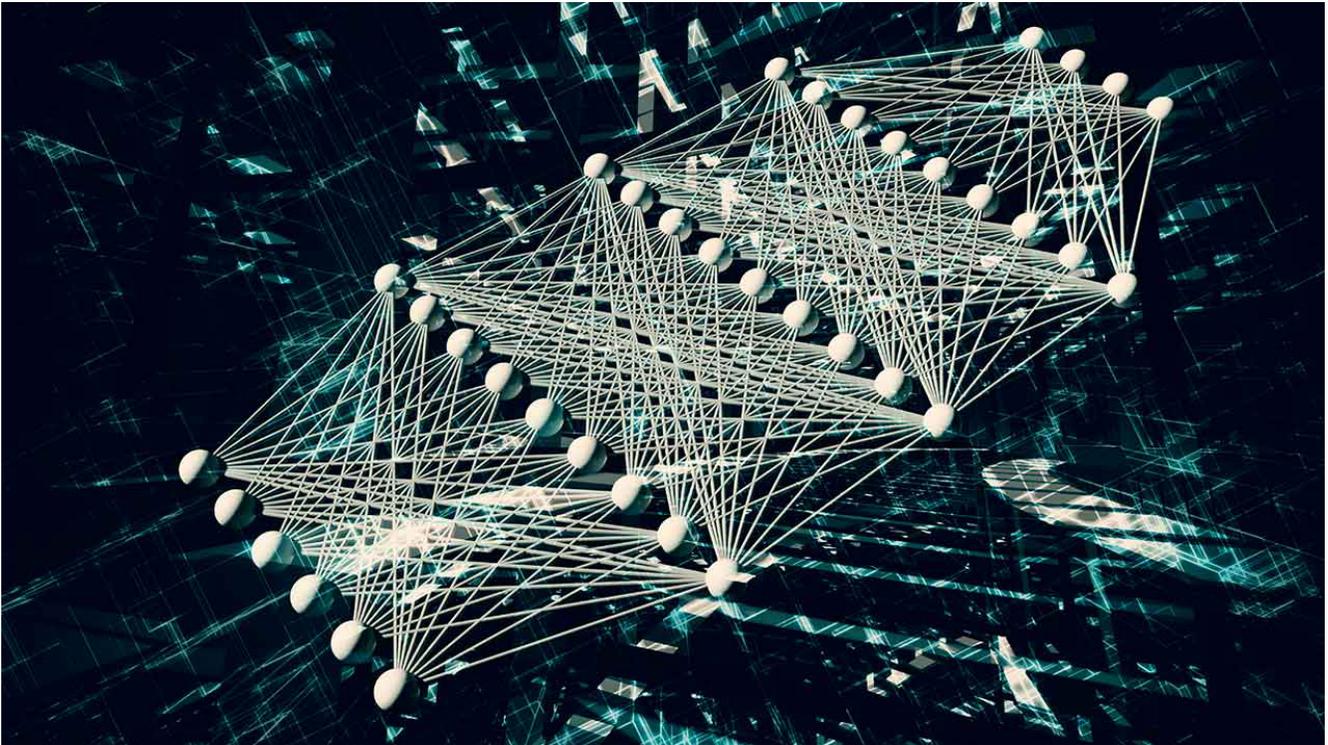
Even if you're not very familiar with deep learning, you've probably heard about it and how it can, among other things, help automate the driving experience, increase manufacturing efficiency and change the consumer shopping experience. Deep learning is the latest evolution of artificial intelligence (AI) and machine learning. Both machine learning and deep learning are subsets of AI. While traditional machine learning algorithms need to be specifically programmed by people with domain-level expertise, deep learning algorithms utilize neural networks that can be trained by feeding them data. The same network can be used to solve very different problems by training it with different data, removing the requirement of having specific expertise in the problem that needs to be solved. Because of this, deep learning is considered a foundational technology that has the potential to make an impact in many industries.

For those new to deep learning, think of it as a technology that can classify things. Many real-world problems can be broken down into a classification problem, for example:

- Autonomous and semi-autonomous vehicles need to be able to classify roads, other vehicles, people and road signs as shown in Figure 1.
- Smart factories need to be able to classify a product as defective
- Automated retail outlets need to classify if a consumer is removing a product from a store for purchase or just looking at it

Figure 1. Classification Example of an Automotive Use Case. Every Object in the Scene Is Determined to Be One of the Five Different Classes.

There are two main parts to deep learning: training and inference, as shown in Figure 2. During training, a model of a neural network that has been created to solve a classification problem is fed huge amounts of labeled data, from which it learns how to classify things. In the example scenario shown in Figure 1 of an autonomous vehicle, the model would be fed images of pedestrians, road signs, cars and roads with each of these objects correctly labeled. Real-time performance or power is not an issue during training, so it usually occurs on a desktop or cloud platform.

During inference, the trained network that has been deployed into its end application makes a classification decision, such as determining whether or not a part on an assembly line is defective.
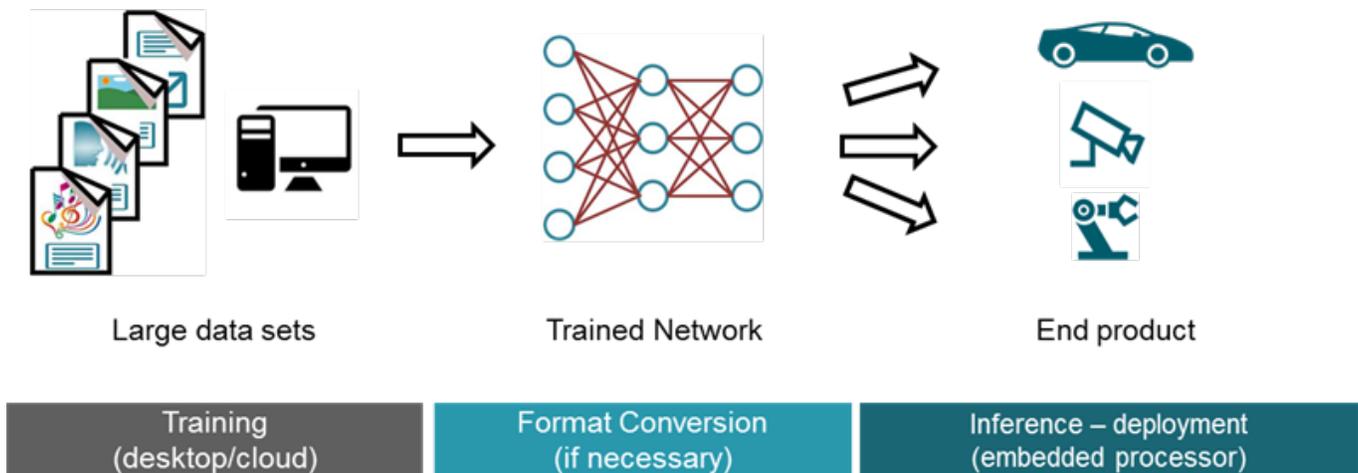

Figure 2. Typical Development Flow of a Deep Learning Network.

For a combination of reasons – including reliability, low latency, privacy, power and cost – it is beneficial for inference to run close to where sensors gather data; that is, at the "edge" of the network. Advanced sensors that can handle some amount of AI are capable of running a simple inference in the sensor itself. For a more complex inference, a separate edge processor is needed.

TI Sitara™ processors are great devices for edge processing. With a scalable portfolio integrating Arm® Cortex®-A cores with flexible peripherals, Sitara processors provide industrial-grade solutions with an extensive mix of processing, integration and power efficiency. With TI Deep Learning (TIDL) software, Sitara AM57x processors leverage hardware acceleration to improve the performance of machine learning and deep learning inference at the edge.

TIDL software is a set of open-source Linux software packages included on the processor software development kit (SDK) and tools that enable the hardware-accelerated offloading of deep learning inference in an AM57x device. It's possible to accelerate inference across embedded vision engine (EVE) subsystems, C66x digital signal processor cores or a combination of both, using a set of application programming interfaces built with Open Computing Language (OpenCL).

For those who already have an existing neural network framework, TIDL software bridges the gap between the popular Caffe and TensorFlow frameworks with accelerated hardware on the AM57x through an import tool, as shown in Figure 3. To maximize efficiency on the hardware, or for those who don't have an existing neural network framework, Caffe Jacinto, a TI-developed framework forked from NVIDIA/Caffe and designed for the embedded space, can perform the training part of deep learning.
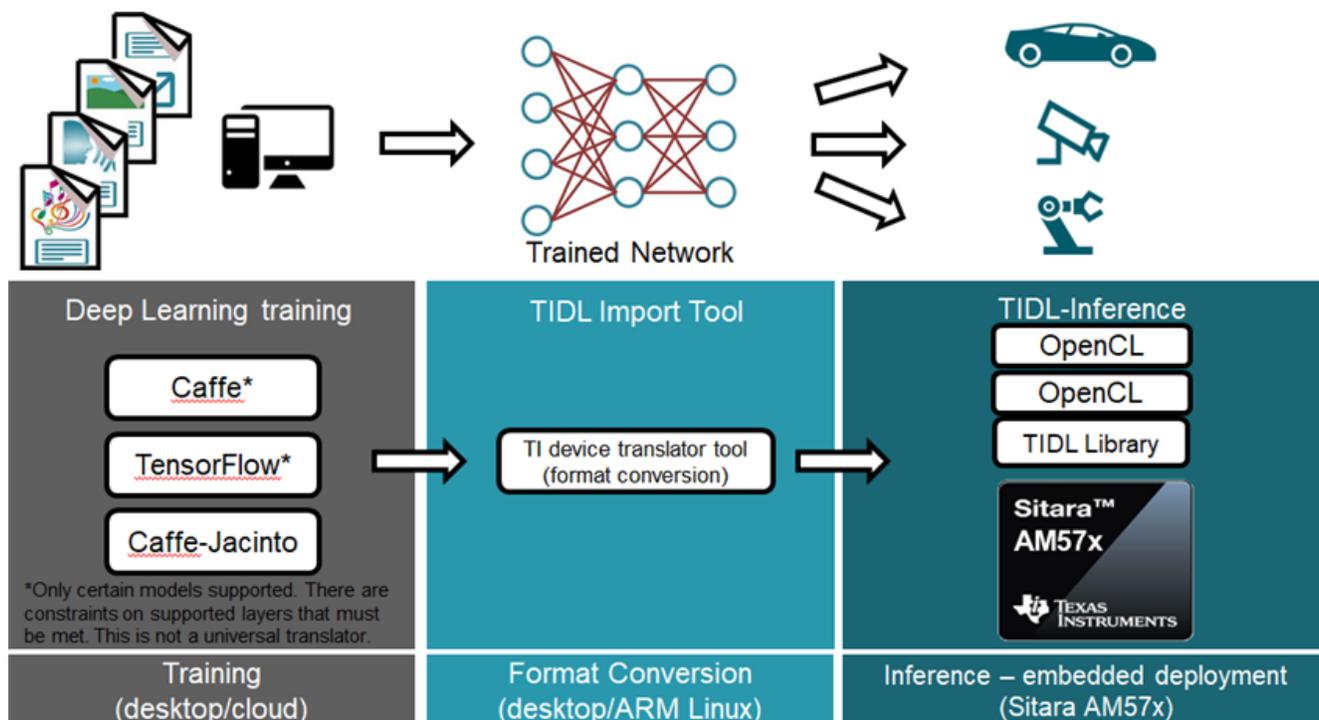
Figure 3. Texas Instruments Deep Learning (TIDL) Software Development Flow.

Note that while TI has developed a framework that can be used to train a network model for an embedded application, the training itself still takes place on a desktop or cloud platform. It is the inference that runs on the TI embedded processor.

With free software as part of TI's Processor SDK, training videos and a deep learning TI reference design, whether you're new to deep learning or a seasoned expert, TI provides the building blocks to get you started quickly and easily.

Additional Resources:
- Visit the TI Sitara Machine Learning page.
- Read more about machine learning:
    – Bringing machine learning to embedded systems
    – AI in Automotive: Practical deep learning
- Download the Deep Learning Inference for Embedded Applications Reference Design
- Watch these TI training videos:
    – TIDL Overview for Sitara Processors

IMPORTANT NOTICE AND DISCLAIMER